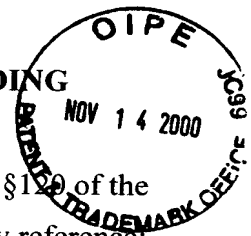


**SEQUENCE-DETERMINED DNA FRAGMENTS AND CORRESPONDING
POLYPEPTIDES ENCODED THEREBY**

This application claims priority under 35 USC §119(e), §119(a-d) and §120 of the following applications, the entire contents of which are hereby incorporated by reference.

Country	Filing Date	Attorney No.	Client No.	Application No.
United States	06/18/99	2750-0465P	00038.001	60/139,763

This application contains a CDR, the entire contents of which are hereby incorporated by reference. The CDR contains the following files:

File Name	Date of Creation	File Size
2750-0942P.ST25	November 13, 2000	102 KB

FIELD OF THE INVENTION

The present invention relates to isolated polynucleotides that represent a complete gene, or a fragment thereof, that is expressed. In addition, the present invention relates to the polypeptide or protein corresponding to the coding sequence of these polynucleotides. The present invention also relates to isolated polynucleotides that represent regulatory regions of genes. The present invention also relates to isolated polynucleotides that represent untranslated regions of genes. The present invention further relates to the use of these isolated polynucleotides and polypeptides and proteins.

DESCRIPTION OF THE RELATED ART

Efforts to map and sequence the genome of a number of organisms are in progress; a few complete genome sequences, for example those of *E. coli* and *Saccharomyces cerevisiae* are known (Blattner et al., *Science* 277:1453 (1997); Goffeau et al., *Science* 274:546 (1996)). The complete genome of a multicellular organism, *C. elegans*, has also been sequenced (See, the *C. elegans* Sequencing Consortium, *Science* 282:2012 (1998)). To date, no complete genome of a plant has been sequenced, nor has a complete cDNA complement of any plant been sequenced.

SUMMARY OF THE INVENTION

The present invention comprises polynucleotides, such as complete cDNA sequences and/or sequences of genomic DNA encompassing complete genes, fragments of genes, and/or regulatory elements of genes and/or regions with other functions and/or intergenic regions,

hereinafter collectively referred to as Sequence-Determined DNA Fragments (SDFs), from different plant species, particularly corn, wheat, soybean, rice and *Arabidopsis thaliana*, and other plants and or mutants, variants, fragments or fusions of said SDFs and polypeptides or proteins derived therefrom. In some instances, the SDFs span the entirety of a protein-coding segment. In some instances, the entirety of an mRNA is represented. Other objects of the invention that are also represented by SDFs of the invention are control sequences, such as, but not limited to, promoters. Complements of any sequence of the invention are also considered part of the invention.

Other objects of the invention are polynucleotides comprising exon sequences, polynucleotides comprising intron sequences, polynucleotides comprising introns together with exons, intron/exon junction sequences, 5' untranslated sequences, and 3' untranslated sequences of the SDFs of the present invention. Polynucleotides representing the joinder of any exons described herein, in any arrangement, for example, to produce a sequence encoding any desirable amino acid sequence are within the scope of the invention.

The present invention also resides in probes useful for isolating and identifying nucleic acids that hybridize to an SDF of the invention. The probes can be of any length, but more typically are 12-2000 nucleotides in length; more typically, 15 to 200 nucleotides long; even more typically, 18 to 100 nucleotides long.

Yet another object of the invention is a method of isolating and/or identifying nucleic acids using the following steps:

- (a) contacting a probe of the instant invention with a polynucleotide sample under conditions that permit hybridization and formation of a polynucleotide duplex; and
- (b) detecting and/or isolating the duplex of step (a).

The conditions for hybridization can be from low to moderate to high stringency conditions. The sample can include a polynucleotide having a sequence unique in a plant genome. Probes and methods of the invention are useful, for example, without limitation, for mapping of genetic traits and/or for positional cloning of a desired fragment of genomic DNA.

Probes and methods of the invention can also be used for detecting alternatively spliced messages within a species. Probes and methods of the invention can further be used to detect or isolate related genes in other plant species using genomic DNA (gDNA) and/or cDNA libraries. In some instances, especially when longer probes and low to moderate stringency hybridization conditions are used; the probe will hybridize to a plurality of cDNA and/or gDNA sequences of a plant. This approach is useful for isolating representatives of gene families which are

identifiable by possession of a common functional domain in the gene product or which have common cis-acting regulatory sequences. This approach is also useful for identifying orthologous genes from other organisms.

The present invention also resides in constructs for modulating the expression of the genes comprised of all or a fragment of an SDF. The constructs comprise all or a fragment of the expressed SDF, or of a complementary sequence. Examples of constructs include ribozymes comprising RNA encoded by an SDF or by a sequence complementary thereto, antisense constructs, constructs comprising coding regions or parts thereof, constructs comprising promoters, introns, untranslated regions, scaffold attachment regions, methylating regions, enhancing or reducing regions, DNA and chromatin conformation modifying sequences, etc. Such constructs can be constructed using viral, plasmid, bacterial artificial chromosomes (BACs), plasmid artificial chromosomes (PACs), autonomous plant plasmids, plant artificial chromosomes or other types of vectors and exist in the plant as autonomous replicating sequences or as DNA integrated into the genome. When inserted into a host cell the construct is, preferably, functionally integrated with, or operatively linked to, a heterologous polynucleotide. For instance, a coding region from an SDF might be operably linked to a promoter that is functional in a plant.

The present invention also resides in host cells, including bacterial or yeast cells or plant cells, and plants that harbor constructs such as described above. Another aspect of the invention relates to methods for modulating expression of specific genes in plants by expression of the coding sequence of the constructs, by regulation of expression of one or more endogenous genes in a plant or by suppression of expression of the polynucleotides of the invention in a plant. Methods of modulation of gene expression include without limitation (1) inserting into a host cell additional copies of a polynucleotide comprising a coding sequence; (2) modulating an endogenous promoter in a host cell; (3) inserting antisense or ribozyme constructs into a host cell and (4) inserting into a host cell a polynucleotide comprising a sequence encoding a variant, fragment, or fusion of the native polypeptides of the instant invention.

BRIEF DESCRIPTION OF THE TABLES

In TABLE 1, the format of the data is as follows:

In Table 1, sequence data are presented in the form of annotation of a reference sequence. The format is shown below. The reference sequence is shown at the top of the annotation file as a 7 digit sequence number preceded by ">" (*e.g.* >5019261). The sequence

identifier is a “gi” number that identifies a specific DNA sequence in the publically accessible BLAST Databases on the NCBI FTP web site (accessible at ncbi.nlm.gov/blast). In particular, the “nt.Z” nucleotide sequence data base at the NCBI FTP site utilizes the “gi” identifiers to assign by NCBI a unique identifier for each sequence in the databases, thereby providing a non-redundant database for sequences from various data bases, including GenBank, EMBL, DDBJ (DNA Database of Japan) and PDB (Brookhaven Protein Data Bank). Thus, the line in TABLE 1 beginning with sequence number identifies the unique “gi” identifier followed by the corresponding GenBank (gb) accession number and locus. The reference sequence number is followed on the next line by data regarding the length of the sequence (“len”) and the number of exons found in the sequence by the analysis program (“nex”).

The annotation data are presented in columns; the leftmost column identifies the position of the putative exon in the gene as initial (“init”), internal (“intr”) or terminal (“term”). Genes considered composed of a single exon are denoted “sngl”. The next column describes the position in the nucleotide sequence beginning the exon (“start”) and the next column describes the position in the nucleotide sequence ending the exon (“stop”). The direction of the gene is indicated in the next column, “+” indicating 5’ - 3’ in the direction presented in the database, “-” indicating the opposite orientation. The “gene number” is given in the final column. Exons having the same gene number are grouped in the order shown to create the relevant coding sequence.

>5019261 ← This is the gi number of the public sequence

len = 97208 nex = 121

↑

↑

Length

Number exons

5 of public sequence

	Exon Type	Start	Stop	Direction	Gene Number
10	↓	↓	↓	↓	↓
	Sngl	602	778	+	0
	Sngl	990	1316	+	1
	Sngl	2356	2691	+	2
	Sngl	4634	4735	+	3
15	Sngl	4973	5092	+	4
	Sngl	5746	5874	+	5
	Init	8119	8798	+	6
	Term	9284	9518	+	6
	Init	10827	11150	+	7
20	Term	11294	11335	+	7
	Sngl	12655	12825	+	8
	Sngl	13303	13596	+	9
	Sngl	18654	18782	+	10
	Sngl	19880	20086	+	11
25	Init	21476	21539	+	12
	Intr	21647	21802	+	12
	Term	23488	23567	+	12
	Init	25035	25133	+	13
	Intr	25466	25589	+	13
30	Intr	25677	25786	+	13
	Intr	25899	25962	+	13
	Intr	26045	26109	+	13
	Intr	26188	26253	+	13
	Term	26350	26448	+	13
35	Sngl	27671	27793	+	14
	Sngl	29126	29299	+	15
	Sngl	30266	30364	+	16
	Sngl	31717	31929	+	17
	Sngl	32102	32209	+	18
40	Sngl	32450	32548	+	19

				6	
	Sngl	32634	32726	+	20
	Init	35603	35743	+	21
	Term	35829	36185	+	21
	Init	36954	37098	+	22
5	Term	38100	38158	+	22
	Init	39635	39944	+	23
	Intr	40242	40372	+	23
	Intr	40462	40695	+	23
	Intr	40815	41070	+	23
10	Intr	41176	41255	+	23
	Intr	42212	42419	+	23
	Intr	42940	43070	+	23
	Intr	43177	43410	+	23
	Intr	43580	43835	+	23
15	Intr	46672	46715	+	23
	Intr	48334	48532	+	23

DETAILED DESCRIPTION OF THE INVENTION

The invention relates to (I) polynucleotides and methods of use thereof, such as

- IA. Probes, Primers and Substrates;
- IB. Methods of Detection and Isolation;
 - B.1. Hybridization;
 - B.2. Methods of Mapping;
 - B.3. Southern Blotting;
 - B.4. Isolating cDNA from Related Organisms;
 - B.5. Isolating and/or Identifying Orthologous Genes
- IC. Methods of Inhibiting Gene Expression
 - C.1. Antisense
 - C.2. Ribozyme Constructs;
 - C.3. Chimeraplasts;
 - C.4. Co-Suppression;
 - C.5. Transcriptional Silencing
 - C.6. Other Methods to Inhibit Gene Expression
- ID. Methods of Functional Analysis;
- IE. Promoter Sequences and Their Use;
- IF. UTRs and/or Intron Sequences and Their Use; and

IG. Coding Sequences and Their Use.

The invention also relates to (II) polypeptides and proteins and methods of use thereof, such as IIA. Native Polypeptides and Proteins

5 A.1 Antibodies

A.2 In Vitro Applications

IIB. Polypeptide Variants, Fragments and Fusions

B.1 Variants

B.2 Fragments

10 B.3 Fusions

The invention also includes (III) methods of modulating polypeptide production, such as

IIIA. Suppression

A.1 Antisense

15 A.2 Ribozymes

A.3 Co-suppression

A.4 Insertion of Sequences into the Gene to be Modulated

A.5 Promoter Modulation

A.6 Expression of Genes containing Dominant-Negative Mutations

20 IIIB. Enhanced Expression

B.1 Insertion of an Exogenous Gene

B.2 Promoter Modulation

The invention further concerns (IV) gene constructs and vector construction, such as

25 IVA. Coding Sequences

IVB. Promoters

IVC. Signal Peptides

The invention still further relates to

30 V Transformation Techniques

Definitions

Allelic variant An “allelic variant” is an alternative form of the same SDF, which resides at the same chromosomal locus in the organism. Allelic variations can occur in any portion of the gene sequence, including regulatory regions. Allelic variants can arise by normal genetic variation in a population. Allelic variants can also be produced by genetic engineering methods. An allelic variant can be one that is found in a naturally occurring plant, including a cultivar or ecotype. An allelic variant may or may not give rise to a phenotypic change, and may or may not be expressed. An allele can result in a detectable change in the phenotype of the trait represented by the locus. A phenotypically silent allele can give rise to a product.

Alternatively spliced messages Within the context of the current invention, “alternatively spliced messages” refers to mature mRNAs originating from a single gene with variations in the number and/or identity of exons, introns and/or intron-exon junctions.

Chimeric The term “chimeric” is used to describe genes, as defined supra, or constructs wherein at least two of the elements of the gene or construct, such as the promoter and the coding sequence and/or other regulatory sequences and/or filler sequences and/or complements thereof, are heterologous to each other.

Constitutive Promoter: Promoters referred to herein as “constitutive promoters” actively promote transcription under most, but not necessarily all, environmental conditions and states of development or cell differentiation. Examples of constitutive promoters include the cauliflower mosaic virus (CaMV) 35S transcript initiation region and the 1’ or 2’ promoter derived from T-DNA of *Agrobacterium tumefaciens*, and other transcription initiation regions from various plant genes, such as the maize ubiquitin-1 promoter, known to those of skill.

Coordinately Expressed: The term “coordinately expressed,” as used in the current invention, refers to genes that are expressed at the same or a similar time and/or stage and/or under the same or similar environmental conditions.

Domain: Domains are fingerprints or signatures that can be used to characterize protein families and/or parts of proteins. Such fingerprints or signatures can comprise conserved (1) primary sequence, (2) secondary structure, and/or (3) three-dimensional conformation. Generally, each domain has been associated with either a family of proteins or motifs. Typically, these families and/or motifs have been correlated with specific *in-vitro* and/or *in-vivo* activities. A domain can be any length, including the entirety of the sequence of a protein. Detailed descriptions of the domains, associated families and motifs, and correlated activities of the polypeptides of the instant invention are described below. Usually, the polypeptides with designated domain(s) can exhibit at least one activity that is exhibited by any polypeptide that comprises the same domain(s).

Endogenous The term “endogenous,” within the context of the current invention refers to any polynucleotide, polypeptide or protein sequence which is a natural part of a cell or organisms regenerated from said cell.

Exogenous “Exogenous,” as referred to within, is any polynucleotide, polypeptide or protein sequence, whether chimeric or not, that is initially or subsequently introduced into the genome of an individual host cell or the organism regenerated from said host cell by any means other than by a sexual cross. Examples of means by which this can be accomplished are described below, and include *Agrobacterium*-mediated transformation (of dicots - *e.g.* Salomon et al. *EMBO J.* 3:141 (1984); Herrera-Estrella et al. *EMBO J.* 2:987 (1983); of monocots, representative papers are those by Escudero et al., *Plant J.* 10:355 (1996), Ishida et al., *Nature Biotechnology* 14:745 (1996), May et al., *Bio/Technology* 13:486 (1995)), biolistic methods (Armaleo et al., *Current Genetics* 17:97 (1990)), electroporation, *in planta* techniques, and the like. Such a plant containing the exogenous nucleic acid is referred to here as a T₀ for the primary transgenic plant and T₁ for the first generation. The term “exogenous” as used herein is also intended to encompass inserting a naturally found element into a non-naturally found location.

Filler sequence: As used herein, “filler sequence” refers to any nucleotide sequence that is inserted into DNA construct to evoke a particular spacing between particular components

such as a promoter and a coding region and may provide an additional attribute such as a restriction enzyme site.

Gene: The term “gene,” as used in the context of the current invention, encompasses all regulatory and coding sequence contiguously associated with a single hereditary unit with a genetic function (see SCHEMATIC 1). Genes can include non-coding sequences that modulate the genetic function that include, but are not limited to, those that specify polyadenylation, transcriptional regulation, DNA conformation, chromatin conformation, extent and position of base methylation and binding sites of proteins that control all of these. Genes comprised of “exons” (coding sequences), which may be interrupted by “introns” (non-coding sequences), encode proteins. A gene’s genetic function may require only RNA expression or protein production, or may only require binding of proteins and/or nucleic acids without associated expression. In certain cases, genes adjacent to one another may share sequence in such a way that one gene will overlap the other. A gene can be found within the genome of an organism, artificial chromosome, plasmid, vector, etc., or as a separate isolated entity.

Gene Family: “Gene family” is used in the current invention to describe a group of functionally related genes, each of which encodes a separate protein.

Heterologous sequences: “Heterologous sequences” are those that are not operatively linked or are not contiguous to each other in nature. For example, a promoter from corn is considered heterologous to an *Arabidopsis* coding region sequence. Also, a promoter from a gene encoding a growth factor from corn is considered heterologous to a sequence encoding the corn receptor for the growth factor. Regulatory element sequences, such as UTRs or 3’ end termination sequences that do not originate in nature from the same gene as the coding sequence originates from, are considered heterologous to said coding sequence. Elements operatively linked in nature and contiguous to each other are not heterologous to each other. On the other hand, these same elements remain operatively linked but become heterologous if other filler sequence is placed between them. Thus, the promoter and coding sequences of a corn gene expressing an amino acid transporter are not heterologous to each other, but the promoter and coding sequence of a corn gene operatively linked in a novel manner are heterologous.

Homologous gene In the current invention, "homologous gene" refers to a gene that shares sequence similarity with the gene of interest. This similarity may be in only a fragment of the sequence and often represents a functional domain such as, examples including without limitation a DNA binding domain, a domain with tyrosine kinase activity, or the like. The functional activities of homologous genes are not necessarily the same.

Inducible Promoter An "inducible promoter" in the context of the current invention refers to a promoter which is regulated under certain conditions, such as light, chemical concentration, protein concentration, conditions in an organism, cell, or organelle, etc. A typical example of an inducible promoter, which can be utilized with the polynucleotides of the present invention, is PARSK1, the promoter from the *Arabidopsis* gene encoding a serine-threonine kinase enzyme, and which promoter is induced by dehydration, abscissic acid and sodium chloride (Wang and Goodman, *Plant J.* 8:37 (1995)) Examples of environmental conditions that may affect transcription by inducible promoters include anaerobic conditions, elevated temperature, or the presence of light.

Intergenic region "Intergenic region," as used in the current invention, refers to nucleotide sequence occurring in the genome that separates adjacent genes.

Mutant gene In the current invention, "mutant" refers to a heritable change in DNA sequence at a specific location. Mutants of the current invention may or may not have an associated identifiable function when the mutant gene is transcribed.

Orthologous Gene In the current invention "orthologous gene" refers to a second gene that encodes a gene product that performs a similar function as the product of a first gene. The orthologous gene may also have a degree of sequence similarity to the first gene. The orthologous gene may encode a polypeptide that exhibits a degree of sequence similarity to a polypeptide corresponding to a first gene. The sequence similarity can be found within a functional domain or along the entire length of the coding sequence of the genes and/or their corresponding polypeptides.

Percentage of sequence identity "Percentage of sequence identity," as used herein, is determined by comparing two optimally aligned sequences over a comparison window, where the fragment of the polynucleotide or amino acid sequence in the comparison window may comprise additions or deletions (e.g., gaps or overhangs) as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. The percentage is calculated by determining the number of positions at which the identical nucleic acid base or amino acid residue occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison and multiplying the result by 100 to yield the percentage of sequence identity. Optimal alignment of sequences for comparison may be conducted by the local homology algorithm of Smith and Waterman *Add. APL. Math.* 2:482 (1981), by the homology alignment algorithm of Needleman and Wunsch *J. Mol. Biol.* 48:443 (1970), by the search for similarity method of Pearson and Lipman *Proc. Natl. Acad. Sci. (USA)* 85: 2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, BLAST, PASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group (GCG), 575 Science Dr., Madison, WI), or by inspection. Given that two sequences have been identified for comparison, GAP and BESTFIT are preferably employed to determine their optimal alignment. Typically, the default values of 5.00 for gap weight and 0.30 for gap weight length are used. The term "substantial sequence identity" between polynucleotide or polypeptide sequences refers to polynucleotide or polypeptide comprising a sequence that has at least 80% sequence identity, preferably at least 85%, more preferably at least 90% and most preferably at least 95%, even more preferably, at least 96%, 97%, 98% or 99% sequence identity compared to a reference sequence using the programs.

Plant Promoter A "plant promoter" is a promoter capable of initiating transcription in plant cells and can drive or facilitate transcription of a fragment of the SDF of the instant invention or a coding sequence of the SDF of the instant invention. Such promoters need not be of plant origin. For example, promoters derived from plant viruses, such as the CaMV35S promoter or from *Agrobacterium tumefaciens* such as the T-DNA promoters, can be plant promoters. A typical example of a plant promoter of plant origin is the maize ubiquitin-1 (ubi-1) promoter known to those of skill.

Promoter: The term "promoter," as used herein, refers to a region of sequence determinants located upstream from the start of transcription of a gene and which are involved in recognition and binding of RNA polymerase and other proteins to initiate and modulate transcription. A basal promoter is the minimal sequence necessary for assembly of a transcription complex required for transcription initiation. Basal promoters frequently include a "TATA box" element usually located between 15 and 35 nucleotides upstream from the site of initiation of transcription. Basal promoters also sometimes include a "CCAAT box" element (typically a sequence CCAAT) and/or a GGGCG sequence, usually located between 40 and 200 nucleotides, preferably 60 to 120 nucleotides, upstream from the start site of transcription.

Public sequence: The term "public sequence," as used in the context of the instant application, refers to any sequence that has been deposited in a publicly accessible database. This term encompasses both amino acid and nucleotide sequences. Such sequences are publicly accessible, for example, on the BLAST databases on the NCBI FTP web site (accessible at ncbi.nlm.gov/blast). The database at the NCBI GTP site utilizes "gi" numbers assigned by NCBI as a unique identifier for each sequence in the databases, thereby providing a non-redundant database for sequence from various databases, including GenBank, EMBL, DDBJ, (DNA Database of Japan) and PDB (Brookhaven Protein Data Bank).

Regulatory Sequence The term "regulatory sequence," as used in the current invention, refers to any nucleotide sequence that influences transcription or translation initiation and rate, and stability and/or mobility of the transcript or polypeptide product. Regulatory sequences include, but are not limited to, promoters, promoter control elements, protein binding sequences, 5' and 3' UTRs, transcriptional start site, termination sequence, polyadenylation sequence, introns, certain sequences within a coding sequence, etc.

Related Sequences: "Related sequences" refer to either a polypeptide or a nucleotide sequence that exhibits some degree of sequence similarity with a sequence described in Table 1.

Scaffold Attachment Region (SAR) As used herein, "scaffold attachment region" is a DNA sequence that anchors chromatin to the nuclear matrix or scaffold to generate loop domains that can have either a transcriptionally active or inactive structure (Spiker and Thompson (1996) *Plant Physiol.* 110: 15-21).

Sequence-determined DNA fragments (SDFs) "Sequence-determined DNA fragments" as used in the current invention are isolated sequences of genes, fragments of genes, intergenic regions or contiguous DNA from plant genomic DNA or cDNA or RNA the sequence of which has been determined.

Signal Peptide A "signal peptide" as used in the current invention is an amino acid sequence that targets the protein for secretion, for transport to an intracellular compartment or organelle or for incorporation into a membrane. Signal peptides are indicated in the tables and a more detailed description located below.

Specific Promoter In the context of the current invention, "specific promoters" refers to a subset of inducible promoters that have a high preference for being induced in a specific tissue or cell and/or at a specific time during development of an organism. By "high preference" is meant at least 3-fold, preferably 5-fold, more preferably at least 10-fold still more preferably at least 20-fold, 50-fold or 100-fold increase in transcription in the desired tissue over the transcription in any other tissue. Typical examples of temporal and/or tissue specific promoters of plant origin that can be used with the polynucleotides of the present invention, are: PTA29, a promoter which is capable of driving gene transcription specifically in tapetum and only during anther development (Koltonow et al., *Plant Cell* 2:1201 (1990); RCc2 and RCc3, promoters that direct root-specific gene transcription in rice (Xu et al., *Plant Mol. Biol.* 27:237 (1995); TobRB27, a root-specific promoter from tobacco (Yamamoto et al., *Plant Cell* 3:371 (1991)). Examples of tissue-specific promoters under developmental control include promoters that initiate transcription only in certain tissues or organs, such as root, ovule, fruit, seeds, or flowers. Other suitable promoters include those from genes encoding storage proteins or the lipid body membrane protein, oleosin. A few root-specific promoters are noted above.

Stringency "Stringency" as used herein is a function of probe length, probe composition (G + C content), and salt concentration, organic solvent concentration, and temperature of

hybridization or wash conditions. Stringency is typically compared by the parameter T_m , which is the temperature at which 50% of the complementary molecules in the hybridization are hybridized, in terms of a temperature differential from T_m . High stringency conditions are those providing a condition of $T_m - 5^\circ\text{C}$ to $T_m - 10^\circ\text{C}$. Medium or moderate stringency conditions are those providing $T_m - 20^\circ\text{C}$ to $T_m - 29^\circ\text{C}$. Low stringency conditions are those providing a condition of $T_m - 40^\circ\text{C}$ to $T_m - 48^\circ\text{C}$. The relationship of hybridization conditions to T_m (in $^\circ\text{C}$) is expressed in the mathematical equation

$$T_m = 81.5 - 16.6(\log_{10}[\text{Na}^+]) + 0.41(\%G+C) - (600/N) \quad (1)$$

where N is the length of the probe. This equation works well for probes 14 to 70 nucleotides in length that are identical to the target sequence. The equation below for T_m of DNA-DNA hybrids is useful for probes in the range of 50 to greater than 500 nucleotides, and for conditions that include an organic solvent (formamide).

$$T_m = 81.5 + 16.6 \log \{ [\text{Na}^+]/(1 + 0.7[\text{Na}^+]) \} + 0.41(\%G+C) - 500/L - 0.63(\%\text{formamide}) \quad (2)$$

where L is the length of the probe in the hybrid. (P. Tijessen, "Hybridization with Nucleic Acid Probes" in Laboratory Techniques in Biochemistry and Molecular Biology, P.C. van der Vliet, ed., c. 1993 by Elsevier, Amsterdam.) The T_m of equation (2) is affected by the nature of the hybrid; for DNA-RNA hybrids T_m is $10\text{-}15^\circ\text{C}$ higher than calculated, for RNA-RNA hybrids T_m is $20\text{-}25^\circ\text{C}$ higher. Because the T_m decreases about 1°C for each 1% decrease in homology when a long probe is used (Bonner et al., *J. Mol. Biol.* **81**:123 (1973)), stringency conditions can be adjusted to favor detection of identical genes or related family members.

Equation (2) is derived assuming equilibrium and therefore, hybridizations according to the present invention are most preferably performed under conditions of probe excess and for sufficient time to achieve equilibrium. The time required to reach equilibrium can be shortened by inclusion of a hybridization accelerator such as dextran sulfate or another high volume polymer in the hybridization buffer.

Stringency can be controlled during the hybridization reaction or after hybridization has occurred by altering the salt and temperature conditions of the wash solutions used. The formulas shown above are equally valid when used to compute the stringency of a wash

solution. Preferred wash solution stringencies lie within the ranges stated above; high stringency is 5-8°C below T_m , medium or moderate stringency is 26-29°C below T_m and low stringency is 45-48°C below T_m .

5 Substantially free of A composition containing A is “substantially free of” B when at least 85% by weight of the total A+B in the composition is A. Preferably, A comprises at least about 90% by weight of the total of A+B in the composition, more preferably at least about 95% or even 99% by weight. For example, a plant gene or DNA sequence can be considered substantially free of other plant genes or DNA sequences.

10

Translational start site In the context of the current invention, a “translational start site” is usually an ATG in the cDNA transcript, more usually the first ATG. A single cDNA, however, may have multiple translational start sites.

15 Transcription start site “Transcription start site” is used in the current invention to describe the point at which transcription is initiated. This point is typically located about 25 nucleotides downstream from a TFIID binding site, such as a TATA box. Transcription can initiate at one or more sites within the gene, and a single gene may have multiple transcriptional start sites, some of which may be specific for transcription in a particular cell-type or tissue.

20

Untranslated region (UTR) A “UTR” is any contiguous series of nucleotide bases that is transcribed, but is not translated. These untranslated regions may be associated with particular functions such as increasing mRNA message stability. Examples of UTRs include, but are not limited to polyadenylation signals, terminations sequences, sequences located
25 between the transcriptional start site and the first exon (5' UTR) and sequences located between the last exon and the end of the mRNA (3' UTR).

Variant: The term “variant” is used herein to denote a polypeptide or protein or polynucleotide molecule that differs from others of its kind in some way. For example,
30 polypeptide and protein variants can consist of changes in amino acid sequence and/or charge and/or post-translational modifications (such as glycosylation, etc).

DETAILED DESCRIPTION OF THE INVENTION

I. Polynucleotides

Exemplified SDFs of the invention represent fragments of the genome of corn, wheat, rice, soybean or *Arabidopsis* and/or represent mRNA expressed from that genome. The isolated nucleic acid of the invention also encompasses corresponding fragments of the genome and/or cDNA complement of other organisms as described in detail below.

Polynucleotides of the invention can be isolated from polynucleotide libraries using primers comprising sequence similar to those described by Table 1. See, for example, the methods described in Sambrook et al., *supra*.

Alternatively, the polynucleotides of the invention can be produced by chemical synthesis. Such synthesis methods are described below.

It is contemplated that the nucleotide sequences presented herein may contain some small percentage of errors. These errors may arise in the normal course of determination of nucleotide sequences. Sequence errors can be corrected by obtaining seeds deposited under the accession numbers cited herein, propagating them, isolating genomic DNA or appropriate mRNA from the resulting plants or seeds thereof, amplifying the relevant fragment of the genomic DNA or mRNA using primers having a sequence that flanks the erroneous sequence, and sequencing the amplification product.

I.A. Probes, Primers and Substrates

SDFs of the invention can be applied to substrates for use in array applications such as, but not limited to, assays of global gene expression, for example under varying conditions of development, growth conditions. The arrays can also be used in diagnostic or forensic methods (WO95/35505, US 5,445,943 and US 5,410,270).

Probes and primers of the instant invention will hybridize to a polynucleotide comprising a sequence in Table 1. Though many different nucleotide sequences can encode an amino acid sequence, the sequences of Table 1 are generally preferred for encoding polypeptides of the invention. However, the sequence of the probes and/or primers of the instant invention need not be identical to those in Table 1 or the complements thereof. For example, some variation in probe or primer sequence and/or length can allow additional

family members to be detected, as well as orthologous genes and more taxonomically distant related sequences. Similarly, probes and/or primers of the invention can include additional nucleotides that serve as a label for detecting the formed duplex or for subsequent cloning purposes.

5 Probe length will vary depending on the application. For use as primers, probes are 12-40 nucleotides, preferably 18-30 nucleotides long. For use in mapping, probes are preferably 50 to 500 nucleotides, preferably 100-250 nucleotides long. For Southern hybridizations, probes as long as several kilobases can be used as explained below.

10 The probes and/or primers can be produced by synthetic procedures such as the triester method of Matteucci et al. *J. Am. Chem. Soc.* 103:3185(1981); or according to Urdea et al. *Proc. Natl. Acad.* 80:7461 (1981) or using commercially available automated oligonucleotide synthesizers.

15 I.B. Methods of Detection and Isolation

The polynucleotides of the invention can be utilized in a number of methods known to those skilled in the art as probes and/or primers to isolate and detect polynucleotides, including, without limitation: Southern, Northern, Branched DNA hybridization assays, polymerase chain reaction, and microarray assays, and variations thereof. Specific methods
20 given by way of examples, and discussed below include:

Hybridization

Methods of Mapping

Southern Blotting

Isolating cDNA from Related Organisms

25 Isolating and/or Identifying Orthologous Genes.

Also, the nucleic acid molecules of the invention can be used in other methods, such as high density oligonucleotide hybridizing assays, described, for example, in U.S. Pat. Nos. 6,004,753; 5,945,306; 5,945,287; 5,945,308; 5,919,686; 5,919,661; 5,919,627; 5,874,248; 5,871,973; 5,871,971; and 5,871,930; and PCT Pub. Nos. WO 9946380; WO 9933981; WO
30 9933870; WO 9931252; WO 9915658; WO 9906572; WO 9858052; WO 9958672; and WO 9810858.

B.1. Hybridization

The isolated SDFs of Table 1 of the present invention can be used as probes and/or primers for detection and/or isolation of related polynucleotide sequences through hybridization. Hybridization of one nucleic acid to another constitutes a physical property that defines the subject SDF of the invention and the identified related sequences. Also, such hybridization imposes structural limitations on the pair. A good general discussion of the factors for determining hybridization conditions is provided by Sambrook et al. ("Molecular Cloning, a Laboratory Manual, 2nd ed., c. 1989 by Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY; *see esp.*, chapters 11 and 12). Additional considerations and details of the physical chemistry of hybridization are provided by G.H. Keller and M.M. Manak "DNA Probes", 2nd Ed. pp. 1-25, c. 1993 by Stockton Press, New York, NY.

Depending on the stringency of the conditions under which these probes and/or primers are used, polynucleotides exhibiting a wide range of similarity to those in Table 1 can be detected or isolated. When the practitioner wishes to examine the result of membrane hybridizations under a variety of stringencies, an efficient way to do so is to perform the hybridization under a low stringency condition, then to wash the hybridization membrane under increasingly stringent conditions.

When using SDFs to identify orthologous genes in other species, the practitioner will preferably adjust the amount of target DNA of each species so that, as nearly as is practical, the same number of genome equivalents are present for each species examined. This prevents faint signals from species having large genomes, and thus small numbers of genome equivalents per mass of DNA, from erroneously being interpreted as absence of the corresponding gene in the genome.

The probes and/or primers of the instant invention can also be used to detect or isolate nucleotides that are "identical" to the probes or primers. Two nucleic acid sequences or polypeptides are said to be "identical" if the sequence of nucleotides or amino acid residues, respectively, in the two sequences is the same when aligned for maximum correspondence as described below.

Isolated polynucleotides within the scope of the invention also include allelic variants of the specific sequences presented in Table 1. The probes and/or primers of the invention can also be used to detect and/or isolate polynucleotides exhibiting at least 80% sequence identity with the sequences of Table 1 or fragments thereof.

With respect to nucleotide sequences, degeneracy of the genetic code provides the possibility to substitute at least one base of the base sequence of a gene with a different base without causing the amino acid sequence of the polypeptide produced from the gene to be changed. Hence, the DNA of the present invention may also have any base sequence that has
5 been changed from a sequence in Table 1 by substitution in accordance with degeneracy of genetic code. References describing codon usage include: Carels *et al.*, *J. Mol. Evol.* 46: 45 (1998) and Fennoy *et al.*, *Nucl. Acids Res.* 21(23): 5294 (1993).

B.2. Mapping

The isolated SDF DNA of the invention can be used to create various types of genetic
10 and physical maps of the genome of corn, Arabidopsis, soybean, rice, wheat, or other plants. Some SDFs may be absolutely associated with particular phenotypic traits, allowing construction of gross genetic maps. While not all SDFs will immediately be associated with a phenotype, all SDFs can be used as probes for identifying polymorphisms associated with phenotypes of interest. Briefly, one method of mapping involves total DNA isolation from
15 individuals. It is subsequently cleaved with one or more restriction enzymes, separated according to mass, transferred to a solid support, hybridized with SDF DNA and the pattern of fragments compared. Polymorphisms associated with a particular SDF are visualized as differences in the size of fragments produced between individual DNA samples after digestion with a particular restriction enzyme and hybridization with the SDF. After
20 identification of polymorphic SDF sequences, linkage studies can be conducted. By using the individuals showing polymorphisms as parents in crossing programs, F2 progeny recombinants or recombinant inbreds, for example, are then analyzed. The order of DNA polymorphisms along the chromosomes can be determined based on the frequency with which they are inherited together versus independently. The closer two polymorphisms are
25 together in a chromosome the higher the probability that they are inherited together. Integration of the relative positions of all the polymorphisms and associated marker SDFs can produce a genetic map of the species, where the distances between markers reflect the recombination frequencies in that chromosome segment.

The use of recombinant inbred lines for such genetic mapping is described for
30 *Arabidopsis* by Alonso-Blanco *et al.* (*Methods in Molecular Biology*, vol.82, “*Arabidopsis Protocols*”, pp. 137-146, J.M. Martinez-Zapater and J. Salinas, eds., c. 1998 by Humana Press, Totowa, NJ) and for corn by Burr (“*Mapping Genes with Recombinant Inbreds*”, pp.

249-254. In Freeling, M. and V. Walbot (Ed.), *The Maize Handbook*, c. 1994 by Springer-Verlag New York, Inc.: New York, NY, USA; Berlin Germany; Burr et al. *Genetics* (1998) 118: 519; Gardiner, J. et al., (1993) *Genetics* 134: 917). This procedure, however, is not limited to plants and can be used for other organisms (such as yeast) or for individual cells.

5 The SDFs of the present invention can also be used for simple sequence repeat (SSR) mapping. Rice SSR mapping is described by Morgante et al. (*The Plant Journal* (1993) 3: 165), Panaud et al. (*Genome* (1995) 38: 1170); Senior et al. (*Crop Science* (1996) 36: 1676), Taramino et al. (*Genome* (1996) 39: 277) and Ahn et al. (*Molecular and General Genetics* (1993) 241: 483-90). SSR mapping can be achieved using various methods. In one instance, polymorphisms are identified when sequence specific probes contained within an SDF flanking an SSR are made and used in polymerase chain reaction (PCR) assays with template DNA from two or more individuals of interest. Here, a change in the number of tandem repeats between the SSR-flanking sequences produces differently sized fragments (U.S. Patent 5,766,847). Alternatively, polymorphisms can be identified by using the PCR
10 fragment produced from the SSR-flanking sequence specific primer reaction as a probe against Southern blots representing different individuals (U.H. Refseth et al., (1997) *Electrophoresis* 18: 1519).

Genetic and physical maps of crop species have many uses. For example, these maps can be used to devise positional cloning strategies for isolating novel genes from the mapped
20 crop species. In addition, because the genomes of closely related species are largely syntenic (that is, they display the same ordering of genes within the genome), these maps can be used to isolate novel alleles from relatives of crop species by positional cloning strategies.

The various types of maps discussed above can be used with the SDFs of the invention to identify Quantitative Trait Loci (QTLs). Many important crop traits, such as the solids content of tomatoes, are quantitative traits and result from the combined interactions of
25 several genes. These genes reside at different loci in the genome, oftentimes on different chromosomes, and generally exhibit multiple alleles at each locus. The SDFs of the invention can be used to identify QTLs and isolate specific alleles as described by de Vicente and Tanksley (*Genetics* 134:585 (1993)). In addition to isolating QTL alleles in present crop
30 species, the SDFs of the invention can also be used to isolate alleles from the corresponding QTL of wild relatives. Transgenic plants having various combinations of QTL alleles can then be created and the effects of the combinations measured. Once a desired allele combination has been identified, crop improvement can be accomplished either through

biotechnological means or by directed conventional breeding programs (for review see Tanksley and McCouch, *Science* 277:1063 (1997)).

In another embodiment, the SDFs can be used to help create physical maps of the genome of corn, *Arabidopsis* and related species. Where SDFs have been ordered on a genetic map, as described above, they can be used as probes to discover which clones in large libraries of plant DNA fragments in YACs, BACs, etc. contain the same SDF or similar sequences, thereby facilitating the assignment of the large DNA fragments to chromosomal positions. Subsequently, the large BACs, YACs, etc. can be ordered unambiguously by more detailed studies of their sequence composition (e.g. Marra et al. (1997) *Genomic Research* 7:1072-1084) and by using their end or other sequences to find the identical sequences in other cloned DNA fragments. The overlapping of DNA sequences in this way allows large contigs of plant sequences to be built that, when sufficiently extended, provide a complete physical map of a chromosome. Sometimes the SDFs themselves will provide the means of joining cloned sequences into a contig.

The patent publication WO95/35505 and U.S. Patents 5,445,943 and 5,410,270 describe scanning multiple alleles of a plurality of loci using hybridization to arrays of oligonucleotides. These techniques are useful for each of the types of mapping discussed above.

Following the procedures described above and using a plurality of the SDFs of the present invention, any individual can be genotyped. These individual genotypes can be used for the identification of particular cultivars, varieties, lines, ecotypes and genetically modified plants or can serve as tools for subsequent genetic studies involving multiple phenotypic traits.

B.3 Southern Blot Hybridization

The sequences from Table 1 can be used as probes for various hybridization techniques. These techniques are useful for detecting target polynucleotides in a sample or for determining whether transgenic plants, seeds or host cells harbor a gene or sequence of interest and thus might be expected to exhibit a particular trait or phenotype.

In addition, the SDFs from the invention can be used to isolate additional members of gene families from the same or different species and/or orthologous genes from the same or different species. This is accomplished by hybridizing an SDF to, for example, a Southern blot containing the appropriate genomic DNA or cDNA. Given the resulting hybridization

data, one of ordinary skill in the art could distinguish and isolate the correct DNA fragments by size, restriction sites, sequence and stated hybridization conditions from a gel or from a library.

Identification and isolation of orthologous genes from closely related species and alleles within a species is particularly desirable because of their potential for crop improvement. Many important crop traits, such as the solid content of tomatoes, result from the combined interactions of the products of several genes residing at different loci in the genome. Generally, alleles at each of these loci can make quantitative differences to the trait. By identifying and isolating numerous alleles for each locus from within or different species, transgenic plants with various combinations of alleles can be created and the effects of the combinations measured. Once a more favorable allele combination has been identified, crop improvement can be accomplished either through biotechnological means or by directed conventional breeding programs (Tanksley et al. *Science* 277:1063(1997)).

The results from hybridizations of the SDFs of the invention to, for example, Southern blots containing DNA from another species can also be used to generate restriction fragment maps for the corresponding genomic regions. These maps provide additional information about the relative positions of restriction sites within fragments, further distinguishing mapped DNA from the remainder of the genome.

Physical maps can be made by digesting genomic DNA with different combinations of restriction enzymes.

Probes for Southern blotting to distinguish individual restriction fragments can range in size from 15 to 20 nucleotides to several thousand nucleotides. More preferably, the probe is 100 to 1,000 nucleotides long for identifying members of a gene family when it is found that repetitive sequences would complicate the hybridization. For identifying an entire corresponding gene in another species, the probe is more preferably the length of the gene, typically 2,000 to 10,000 nucleotides, but probes 50-1,000 nucleotides long might be used. Some genes, however, might require probes up to 1,500 nucleotides long or overlapping probes constituting the full-length sequence to span their lengths.

Also, while it is preferred that the probe be homogeneous with respect to its sequence, it is not necessary. For example, as described below, a probe representing members of a gene family having diverse sequences can be generated using PCR to amplify genomic DNA or

RNA templates using primers derived from SDFs that include sequences that define the gene family.

For identifying corresponding genes in another species, the next most preferable probe is a cDNA spanning the entire coding sequence, which allows all of the mRNA-coding fragment of the gene to be identified. Probes for Southern blotting can easily be generated from SDFs by making primers having the sequence at the ends of the SDF and using corn or *Arabidopsis* genomic DNA as a template. In instances where the SDF includes sequence conserved among species, primers including the conserved sequence can be used for PCR with genomic DNA from a species of interest to obtain a probe.

Similarly, if the SDF includes a domain of interest, that fragment of the SDF can be used to make primers and, with appropriate template DNA, used to make a probe to identify genes containing the domain. Alternatively, the PCR products can be resolved, for example by gel electrophoresis, and cloned and/or sequenced. Using Southern hybridization, the variants of the domain among members of a gene family, both within and across species, can be examined.

B.4.1 Isolating DNA from Related Organisms

The SDFs of the invention can be used to isolate the corresponding DNA from other organisms. Either cDNA or genomic DNA can be isolated. For isolating genomic DNA, a lambda, cosmid, BAC or YAC, or other large insert genomic library from the plant of interest can be constructed using standard molecular biology techniques as described in detail by Sambrook et al. 1989 (Molecular Cloning: A Laboratory Manual, 2nd ed. Cold Spring Harbor Laboratory Press, New York) and by Ausubel et al. 1992 (Current Protocols in Molecular Biology, Greene Publishing, New York).

To screen a phage library, for example, recombinant lambda clones are plated out on appropriate bacterial medium using an appropriate *E. coli* host strain. The resulting plaques are lifted from the plates using nylon or nitrocellulose filters. The plaque lifts are processed through denaturation, neutralization, and washing treatments following the standard protocols outlined by Ausubel et al. (1992). The plaque lifts are hybridized to either radioactively labeled or non-radioactively labeled SDF DNA at room temperature for about 16 hours, usually in the presence of 50% formamide and 5X SSC (sodium chloride and sodium citrate) buffer and blocking reagents. The plaque lifts are then washed at 42°C with 1% Sodium Dodecyl Sulfate (SDS) and at a particular concentration of SSC. The SSC concentration used

is dependent upon the stringency at which hybridization occurred in the initial Southern blot analysis performed. For example, if a fragment hybridized under medium stringency (e.g., $T_m - 20^\circ\text{C}$), then this condition is maintained or preferably adjusted to a less stringent condition (e.g., $T_m - 30^\circ\text{C}$) to wash the plaque lifts. Positive clones show detectable hybridization e.g., by exposure to X-ray films or chromogen formation. The positive clones are then subsequently isolated for purification using the same general protocol outlined above. Once the clone is purified, restriction analysis can be conducted to narrow the region corresponding to the gene of interest. The restriction analysis and succeeding subcloning steps can be done using procedures described by, for example Sambrook et al. (1989) cited above.

The procedures outlined for the lambda library are essentially similar to those used for YAC library screening, except that the YAC clones are harbored in bacterial colonies. The YAC clones are plated out at reasonable density on nitrocellulose or nylon filters supported by appropriate bacterial medium in petri plates. Following the growth of the bacterial clones, the filters are processed through the denaturation, neutralization, and washing steps following the procedures of Ausubel et al. 1992. The same hybridization procedures for lambda library screening are followed.

To isolate cDNA, similar procedures using appropriately modified vectors are employed. For instance, the library can be constructed in a lambda vector appropriate for cloning cDNA such as $\lambda\text{gt}11$. Alternatively, the cDNA library can be made in a plasmid vector. cDNA for cloning can be prepared by any of the methods known in the art, but is preferably prepared as described above. Preferably, a cDNA library will include a high proportion of full-length clones.

B. 5. Isolating and/or Identifying Orthologous Genes

Probes and primers of the invention can be used to identify and/or isolate polynucleotides related to those in Table 1. Related polynucleotides are those that are native to other plant organisms and exhibit either similar sequence or encode polypeptides with similar biological activity. One specific example is an orthologous gene. Orthologous genes have the same functional activity. As such, orthologous genes may be distinguished from homologous genes. The percentage of identity is a function of evolutionary separation and, in closely related species, the percentage of identity can be 98 to 100%. The amino acid sequence of a protein encoded by an orthologous gene can be less than 75% identical, but tends to be at least 75% or at

least 80% identical, more preferably at least 90%, most preferably at least 95% identical to the amino acid sequence of the reference protein.

To find orthologous genes, the probes are hybridized to nucleic acids from a species of interest under low stringency conditions, preferably one where sequences containing as much as 40-45% mismatches will be able to hybridize. This condition is established by $T_m - 40^\circ\text{C}$ to $T_m - 48^\circ\text{C}$ (*see* below). Blots are then washed under conditions of increasing stringency. It is preferable that the wash stringency be such that sequences that are 85 to 100% identical will hybridize. More preferably, sequences 90 to 100% identical will hybridize and most preferably only sequences greater than 95% identical will hybridize. One of ordinary skill in the art will recognize that, due to degeneracy in the genetic code, amino acid sequences that are identical can be encoded by DNA sequences as little as 67% identical or less. Thus, it is preferable, for example, to make an overlapping series of shorter probes, on the order of 24 to 45 nucleotides, and individually hybridize them to the same arrayed library to avoid the problem of degeneracy introducing large numbers of mismatches.

As evolutionary divergence increases, genome sequences also tend to diverge. Thus, one of skill will recognize that searches for orthologous genes between more divergent species will require the use of lower stringency conditions compared to searches between closely related species. Also, degeneracy of the genetic code is more of a problem for searches in the genome of a species more distant evolutionarily from the species that is the source of the SDF probe sequences.

Therefore the method described in Bouckaert et al., U.S. Ser. No. 60/121,700 Atty. Dkt. No. 2750-117P, Client Dkt. No. 00010.001, filed February 25, 1999, hereby incorporated in its entirety by reference, can be applied to the SDFs of the present invention to isolate related genes from plant species which do not hybridize to the corn *Arabidopsis*, soybean, rice, wheat, and other plant sequences of Table 1.

Identification of the relationship of nucleotide or amino acid sequences among plant species can be done by comparing the nucleotide or amino acid sequences of SDFs of the present application with nucleotide or amino acid sequences of other SDFs such as those present in applications listed in the table below:

Attorney Docket	Client Docket	Filing Date	Application
2750-0301P	80002.001	9/4/1998	60/099,672
2750-0300P	80001.001	9/4/1998	60/099,671
2750-0302P	80003.001	9/11/1998	60/099,933
2750-0304P	80004.001	9/17/1998	60/100,864
2750-0305P	80005.001	9/18/1998	60/101,042

Attorney Docket	Client Docket	Filing Date	Application
2750-0306P	80006.001	9/21/1998	60/101,255
2750-0307P	80007.001	9/24/1998	60/101,682
2750-0308P	80008.001	9/30/1998	60/102,533
2750-0309P	80009.001	9/30/1998	60/102,460
2750-0310P	80010.001	10/5/1998	60/103,116
2750-0311P	80011.001	10/5/1998	60/103,141
2750-0312P	80012.001	10/6/1998	60/103,215
2750-0313P	80013.001	10/8/1998	60/103,554
2750-0314P	80014.001	10/9/1998	60/103,574
2750-0315P	80015.001	10/13/1998	60/103,907
2750-0316P	80016.001	10/14/1998	60/104,268
2750-0317P	80017.001	10/16/1998	60/104,680
2750-0318P	80018.001	10/19/1998	60/104,828
2750-0319P	80019.001	10/20/1998	60/105,008
2750-0320P	80020.001	10/21/1998	60/105,142
2750-0321P	80021.001	10/22/1998	60/105,533
2750-0322P	80022.001	10/26/1998	60/105,571
2750-0323P	80023.001	10/27/1998	60/105,815
2750-0324P	80024.001	10/29/1998	60/106,105
2750-0325P	80025.001	10/30/1998	60/106,218
2750-0326P	80026.001	11/2/1998	60/106,685
2750-0327P	80027.001	11/6/1998	60/107,282
2750-0329P	80029.001	11/9/1998	60/107,719
2750-0328P	80028.001	11/9/1998	60/107,720
2750-0330P	80030.001	11/10/1998	60/107,836
2750-0331P	80031.001	11/12/1998	60/108,190
2750-0332P	80032.001	11/16/1998	60/108,526
2750-0333P	80033.001	11/17/1998	60/108,901
2750-0335P	80035.001	11/19/1998	60/109,127
2750-0334P	80034.001	11/19/1998	60/109,124
2750-0336P	80036.001	11/20/1998	60/109,267
2750-0337P	80037.001	11/23/1998	60/109,594
2750-0338P	80038.001	11/25/1998	60/110,053
2750-0339P	80039.001	11/25/1998	60/110,050
2750-0340P	80040.001	11/27/1998	60/110,158
2750-0341P	80041.001	11/30/1998	60/110,263
2750-0342P	80042.001	12/1/1998	60/110,495
2750-0343P	80043.001	12/2/1998	60/110,626
2750-0344P	80044.001	12/3/1998	60/110,701
2750-0345P	80045.001	12/7/1998	60/111,339
2750-0346P	80046.001	12/9/1998	60/111,589
2750-0347P	80047.001	12/10/1998	60/111,782
2750-0348P	80048.001	12/11/1998	60/111,812
2750-0349P	80049.001	12/14/1998	60/112,096
2750-0350P	80050.001	12/15/1998	60/112,224
2750-0351P	80051.001	12/16/1998	60/112,624
2750-0352P	80052.001	12/17/1998	60/112,862
2750-0353P	80053.001	12/18/1998	60/112,912
2750-0354P	80054.001	12/21/1998	60/113,248

Attorney Docket	Client Docket	Filing Date	Application
2750-0355P	80055.001	12/22/1998	60/113,522
2750-0356P	80056.001	12/23/1998	60/113,826
2750-0357P	80057.001	12/28/1998	60/113,998
2750-0358P	80058.001	12/29/1998	60/114,384
2750-0359P	80059.001	12/30/1998	60/114,455
2750-0360P	80060.001	1/4/1999	60/114,740
2750-0361P	80061.001	1/6/1999	60/114,866
2750-0362P	80062.001	1/7/1999	60/115,153
2750-0367P	80067.001	1/7/1999	60/115,154
2750-0366P	80066.001	1/7/1999	60/115,156
2750-0365P	80065.001	1/7/1999	60/115,155
2750-0363P	80063.001	1/7/1999	60/115,152
2750-0364P	80064.001	1/7/1999	60/115,151
2750-0370P	80070.001	1/8/1999	60/115,293
2750-0369P	80069.001	1/8/1999	60/115,365
2750-0368P	80068.001	1/8/1999	60/115,364
2750-0371P	80071.001	1/11/1999	60/115,339
2750-0372P	80072.001	1/12/1999	60/115,518
2750-0373P	80073.001	1/13/1999	60/115,847
2750-0374P	80074.001	1/14/1999	60/115,905
2750-0375P	80075.001	1/15/1999	60/116,383
2750-0376P	80076.001	1/15/1999	60/116,384
2750-0378P	80078.001	1/19/1999	60/116,340
2750-0377P	80077.001	1/19/1999	60/116,329
2750-0380P	80080.001	1/21/1999	60/116,672
2750-0379P	80079.001	1/21/1999	60/116,674
2750-0381P	80081.001	1/22/1999	60/116,960
2750-0382P	80082.001	1/22/1999	60/116,962
2750-0383P	80083.001	1/28/1999	60/117,756
2750-0384P	80084.001	2/3/1999	60/118,672
2750-0385P	80085.001	2/4/1999	60/118,808
2750-0386P	80086.001	2/5/1999	60/118,778
2750-0387P	80087.001	2/8/1999	60/119,029
2750-0388P	80088.001	2/9/1999	60/119,332
2750-0389P	80089.001	2/10/1999	60/119,462
2750-0391P	80091.001	2/12/1999	60/119,922
2750-0392P	80092.001	2/16/1999	60/120,196
2750-0393P	80093.001	2/16/1999	60/120,198
2750-0394P	80094.001	2/18/1999	60/120,583
2750-0395P	80095.001	2/22/1999	60/121,072
2750-0396P	80096.001	2/23/1999	60/121,334
2750-0397P	80097.001	2/24/1999	60/121,470
2750-0390P	80090.001	2/25/1999	60/121,825
2750-0398P	80098.001	2/25/1999	60/121,704
2750-0399P	80099.001	2/26/1999	60/122,107
2750-0400P	80100.001	3/1/1999	60/122,266
2750-0401P	80101.001	3/2/1999	60/122,568
2750-0402P	80102.001	3/3/1999	60/122,611
2750-0403P	80103.001	3/4/1999	60/121,775

Attorney Docket	Client Docket	Filing Date	Application
2750-0405P	80105.001	3/5/1999	60/123,180
2750-0404P	80104.001	3/5/1999	60/123,534
2750-0406P	80106.001	3/9/1999	60/123,680
2750-0407P	80107.001	3/9/1999	60/123,548
2750-0408P	80108.001	3/10/1999	60/123,715
2750-0409P	80109.001	3/10/1999	60/123,726
2750-0410P	80110.001	3/11/1999	60/124,263
2750-0411P	80111.001	3/12/1999	60/123,941
2750-0412P	80112.001	3/23/1999	60/125,788
2750-0413P	80113.001	3/25/1999	60/126,264
2750-0414P	80114.001	3/29/1999	60/126,785
2750-0415P	80115.001	4/1/1999	60/127,462
2750-0416P	91000.001	4/6/1999	60/128,234
2750-0417P	91001.001	4/8/1999	60/128,714
2750-0418P	80118.001	4/16/1999	60/129,845
2750-0420P	80120.001	4/19/1999	60/130,077
2750-0421P	80121.001	4/21/1999	60/130,449
2750-0303P	80115.002	4/23/1999	60/130,510
2750-0422P	80122.001	4/23/1999	60/130,891
2750-0423P	80123.001	4/28/1999	60/131,449
2750-0424P	80124.001	4/30/1999	60/132,407
2750-0425P	80125.001	4/30/1999	60/132,048
2750-0426P	80126.001	5/4/1999	60/132,484
2750-0427P	80127.001	5/5/1999	60/132,485
2750-0428P	91002.001	5/6/1999	60/132,487
2750-0429P	80129.001	5/6/1999	60/132,486
2750-0430P	80130.001	5/7/1999	60/132,863
2750-0431P	80131.001	5/11/1999	60/134,256
2750-0433P	00025.001	5/14/1999	60/134,221
2750-0432P	91006.001	5/14/1999	60/134,370
2750-0434P	80116.001	5/14/1999	60/134,219
2750-0435P	80117.001	5/14/1999	60/134,218
2750-0436P	91007.001	5/18/1999	60/134,768
2750-0437P	91008.001	5/19/1999	60/134,941
2750-0438P	91009.001	5/20/1999	60/135,124
2750-0439P	91010.001	5/21/1999	60/135,353
2750-0440P	91011.001	5/24/1999	60/135,629
2750-0441P	91012.001	5/25/1999	60/136,021
2750-0442P	91013.001	5/27/1999	60/136,392
2750-0444P	91014.001	5/28/1999	60/136,782
2750-0445P	91015.001	6/1/1999	60/137,222
2750-0446P	91016.001	6/3/1999	60/137,528
2750-0447P	91017.001	6/4/1999	60/137,502
2750-0449P	91018.001	6/7/1999	60/137,724
2750-0450P	91019.001	6/8/1999	60/138,094
2750-0457P	00033.001	6/10/1999	60/138,540
2750-0458P	00033.002	6/10/1999	60/138,847
2750-0463P	00034.001	6/14/1999	60/139,119
2750-0461P	80132.011	6/16/1999	60/139,453

Attorney Docket	Client Docket	Filing Date	Application
2750-0462P	80132.012	6/16/1999	60/139,452
2750-0464P	00037.001	6/17/1999	60/139,492
2750-0453P	80132.005	6/18/1999	60/139,462
2750-0466P	00039.001	6/18/1999	60/139,750
2750-0465P	00038.001	6/18/1999	60/139,763
2750-0460P	80132.010	6/18/1999	60/139,455
2750-0451P	80132.003	6/18/1999	60/139,459
2750-0454P	80132.006	6/18/1999	60/139,457
2750-0459P	80132.009	6/18/1999	60/139,463
2750-0448P	80132.002	6/18/1999	60/139,454
2750-0443P	80132.001	6/18/1999	60/139,458
2750-0456P	80132.008	6/18/1999	60/139,456
2750-0455P	80132.007	6/18/1999	60/139,460
2750-0452P	80132.004	6/18/1999	60/139,461
2750-0467P	00042.001	6/21/1999	60/139,817
2750-0468P	00043.001	6/22/1999	60/139,899
2750-0470P	00042.002	6/23/1999	60/140,353
2750-0469P	00044.001	6/23/1999	60/140,354
2750-0471P	00045.001	6/24/1999	60/140,695
2750-0472P	00046.001	6/28/1999	60/140,823
2750-0473P	00048.001	6/29/1999	60/140,991
2750-0474P	00049.001	6/30/1999	60/141,287
2750-0475P	00050.001	7/1/1999	60/141,842
2750-0476P	00051.001	7/1/1999	60/142,154
2750-0477P	00052.001	7/2/1999	60/142,055
2750-0478P	00053.001	7/6/1999	60/142,390
2750-0479P	00054.001	7/8/1999	60/142,803
2750-0480P	00058.001	7/9/1999	60/142,920
2750-0481P	00059.001	7/12/1999	60/142,977
2750-0482P	00060.001	7/13/1999	60/143,542
2750-0489P	00061.001	7/14/1999	60/143,624
2750-0490P	00062.001	7/15/1999	60/144,005
2750-0485P	80134.003	7/16/1999	60/144,086
2750-0486P	80134.004	7/16/1999	60/144,085
2750-0497P	00064.001	7/19/1999	60/144,325
2750-0496P	80134.014	7/19/1999	60/144,334
2750-0495P	80134.013	7/19/1999	60/144,335
2750-0494P	80134.010	7/19/1999	60/144,333
2750-0492P	80134.008	7/19/1999	60/144,331
2750-0488P	80134.006	7/19/1999	60/144,332
2750-0500P	00065.001	7/20/1999	60/144,632
2750-0502P	80135.002	7/20/1999	60/144,884
2750-0499P	80134.012	7/20/1999	60/144,352
2750-0503P	00066.001	7/21/1999	60/144,814
2750-0483P	80134.001	7/21/1999	60/145,088
2750-0484P	80134.002	7/21/1999	60/145,086
2750-0504P	00067.001	7/22/1999	60/145,192
2750-0491P	80134.007	7/22/1999	60/145,085
2750-0493P	80134.009	7/22/1999	60/145,087

Attorney Docket	Client Docket	Filing Date	Application
2750-0487P	80134.005	7/22/1999	60/145,089
2750-0498P	80134.011	7/23/1999	60/145,145
2750-0501P	80135.001	7/23/1999	60/145,224
2750-0505P	00069.001	7/23/1999	60/145,218
2750-0506P	00070.001	7/26/1999	60/145,276
2750-0507P	80136.001	7/27/1999	60/145,918
2750-0508P	80136.002	7/27/1999	60/145,919
2750-0509P	00071.001	7/27/1999	60/145,913
2750-0510P	00072.001	7/28/1999	60/145,951
2750-0513P	00073.001	8/2/1999	60/146,386
2750-0512P	80137.002	8/2/1999	60/146,389
2750-0511P	80137.001	8/2/1999	60/146,388
2750-0514P	00074.001	8/3/1999	60/147,038
2750-0515P	00076.001	8/4/1999	60/147,204
2750-0517P	80138.002	8/4/1999	60/147,302
2750-0519P	80136.003	8/5/1999	60/147,192
2750-0518P	00077.001	8/5/1999	60/147,260
2750-0520P	00079.001	8/6/1999	60/147,416
2750-0516P	80138.001	8/6/1999	60/147,303
2750-0523P	80139.002	8/9/1999	60/147,935
2750-0521P	00080.001	8/9/1999	60/147,493
2750-0522P	80139.001	8/10/1999	60/148,171
2750-0524P	00081.001	8/11/1999	60/148,319
2750-0530P	00082.001	8/12/1999	60/148,341
2750-0525P	80141.001	8/12/1999	60/148,347
2750-0526P	80141.002	8/12/1999	60/148,342
2750-0527P	80141.003	8/12/1999	60/148,340
2750-0528P	80141.004	8/12/1999	60/148,337
2750-0532P	80142.002	8/13/1999	60/148,684
2750-0529P	00083.001	8/13/1999	60/148,565
2750-0531P	80142.001	8/16/1999	60/149,368
2750-0533P	80001.002	8/17/1999	60/149,927
2750-0534P	80001.003	8/17/1999	60/149,928
2750-0535P	80001.004	8/17/1999	60/149,926
2750-0536P	80001.005	8/17/1999	60/149,925
2750-0537P	00084.001	8/17/1999	60/149,175
2750-0538P	00085.001	8/18/1999	60/149,426
2750-0542P	00087.001	8/20/1999	60/149,723
2750-0541P	80143.002	8/20/1999	60/149,929
2750-0539P	00086.001	8/20/1999	60/149,722
2750-0543P	00088.001	8/23/1999	60/149,902
2750-0540P	80143.001	8/23/1999	60/149,930
2750-0544P	00089.001	8/25/1999	60/150,566
2750-0547P	00090.001	8/26/1999	60/150,884
2750-0546P	80144.002	8/27/1999	60/151,066
2750-0548P	00091.001	8/27/1999	60/151,080
2750-0545P	80144.001	8/27/1999	60/151,065
2750-0549P	00092.001	8/30/1999	60/151,303
2750-0552P	00093.001	8/31/1999	60/151,438

Attorney Docket	Client Docket	Filing Date	Application
2750-0553P	00094.001	9/1/1999	60/151,930
2750-0550P	80001.006	9/3/1999	09/391,631
2750-0551F(PC)	80001.100	9/3/1999	99/204,38
2750-0554P	00095.001	9/7/1999	60/152,363
2750-0555P	00096.001	9/10/1999	60/153,070
2750-0556P	00098.001	9/13/1999	60/153,758
2750-0557P	00099.001	9/15/1999	60/154,018
2750-0558P	00101.001	9/16/1999	60/154,039
2750-0559P	00102.001	9/20/1999	60/154,779
2750-0560P	00103.001	9/22/1999	60/155,139
2750-0561P	00104.001	9/23/1999	60/155,486
2750-0562P	00105.001	9/24/1999	60/155,659
2750-0563P	00106.001	9/28/1999	60/156,458
2750-0564P	00107.001	9/29/1999	60/156,596
2750-0570P	00108.001	10/4/1999	60/157,117
2750-0571P	00109.001	10/5/1999	60/157,753
2750-0565P	80010.002	10/5/1999	09/413,198
2750-0566P	80010.003	10/5/1999	09/412,922
2750-0567F(PC)	80010.100	10/5/1999	99/228,55
2750-0568F(PC)	80010.101	10/5/1999	99/228,54
2750-0569F(PC)	80010.102	10/5/1999	99/228,53
2750-0572P	00110.001	10/6/1999	60/157,865
2750-0575P	00111.001	10/7/1999	60/158,029
2750-0576P	00112.001	10/8/1999	60/158,232
2750-0577P	00113.001	10/12/1999	60/158,369
2750-0574P	80145.002	10/13/1999	60/159,295
2750-0579P	80146.002	10/13/1999	60/159,293
2750-0583P	80148.002	10/13/1999	60/159,294
2750-0573P	80145.001	10/14/1999	60/159,330
2750-0580P	80147.001	10/14/1999	60/159,638
2750-0581P	80147.002	10/14/1999	60/159,637
2750-0582P	80148.001	10/14/1999	60/159,329
2750-0578P	80146.001	10/14/1999	60/159,331
2750-0584P	00116.001	10/18/1999	60/159,584
2750-0586P	80149.001	10/21/1999	60/160,814
2750-0587P	80149.002	10/21/1999	60/160,770
2750-0588P	00119.001	10/21/1999	60/160,741
2750-0589P	80150.001	10/21/1999	60/160,768
2750-0590P	80150.002	10/21/1999	60/160,767
2750-0585P	00118.001	10/21/1999	60/160,815
2750-0593P	80151.002	10/22/1999	60/160,981
2750-0591P	00120.001	10/22/1999	60/160,980
2750-0592P	80151.001	10/22/1999	60/160,989
2750-0596P	80152.002	10/25/1999	60/161,404
2750-0595P	80152.001	10/25/1999	60/161,406
2750-0594P	00121.001	10/25/1999	60/161,405
2750-0597P	00122.001	10/26/1999	60/161,361
2750-0598P	80153.001	10/26/1999	60/161,360
2750-0599P	80153.002	10/26/1999	60/161,359

Attorney Docket	Client Docket	Filing Date	Application
2750-0600P	80026.002	10/28/1999	09/428,944
2750-0601P	00123.001	10/28/1999	60/161,920
2750-0602P	80154.001	10/28/1999	60/161,992
2750-0603P	80154.002	10/28/1999	60/161,993
2750-0604P	00124.001	10/29/1999	60/162,143
2750-0605P	80155.001	10/29/1999	60/162,142
2750-0606P	80155.002	10/29/1999	60/162,228
2750-0607P	00125.001	11/1/1999	60/162,894
2750-0608P	80156.001	11/1/1999	60/162,891
2750-0609P	80156.002	11/1/1999	60/162,895
2750-0610P	00126.001	11/2/1999	60/163,093
2750-0611P	80157.001	11/2/1999	60/163,092
2750-0612P	80157.002	11/2/1999	60/163,091
2750-0614P	80158.001	11/3/1999	60/163,248
2750-0615P	80158.002	11/3/1999	60/163,281
2750-0613P	00127.001	11/3/1999	60/163,249
2750-0618P	80159.002	11/4/1999	60/163,380
2750-0617P	80159.001	11/4/1999	60/163,381
2750-0616P	00128.001	11/4/1999	60/163,379
2750-0621P	80160.002	11/8/1999	60/164,150
2750-0620P	80160.001	11/8/1999	60/164,151
2750-0619P	00129.001	11/8/1999	60/164,146
2750-0623P	80161.002	11/9/1999	60/164,260
2750-0625P	80162.002	11/9/1999	60/164,259
2750-0626P	80163.001	11/10/1999	60/164,321
2750-0630P	80164.002	11/10/1999	60/164,548
2750-0629P	80164.001	11/10/1999	60/164,545
2750-0627P	80163.002	11/10/1999	60/164,318
2750-0624P	80162.001	11/10/1999	60/164,317
2750-0622P	80161.001	11/10/1999	60/164,319
2750-0628P	00131.001	11/10/1999	60/164,544
2750-0636P	80166.002	11/12/1999	60/164,962
2750-0633P	80165.002	11/12/1999	60/164,960
2750-0634P	00133.001	11/12/1999	60/164,870
2750-0632P	80165.001	11/12/1999	60/164,871
2750-0631P	00132.001	11/12/1999	60/164,961
2750-0635P	80166.001	11/12/1999	60/164,959
2750-0637P	00134.001	11/15/1999	60/164,927
2750-0638P	80167.001	11/15/1999	60/164,929
2750-0639P	80167.002	11/15/1999	60/164,926
2750-0640P	00135.001	11/16/1999	60/165,669
2750-0642P	80168.002	11/16/1999	60/165,661
2750-0641P	80168.001	11/16/1999	60/165,671
2750-0643P	00136.001	11/17/1999	60/165,919
2750-0644P	80169.001	11/17/1999	60/165,918
2750-0645P	80169.002	11/17/1999	60/165,911
2750-0646P	00137.001	11/18/1999	60/166,157
2750-0647P	80170.001	11/18/1999	60/166,173
2750-0648P	80170.002	11/18/1999	60/166,158

Attorney Docket	Client Docket	Filing Date	Application
2750-0650P	80171.001	11/19/1999	60/166,411
2750-0649P	00139.001	11/19/1999	60/166,419
2750-0651P	80171.002	11/19/1999	60/166,412
2750-0653P	80172.001	11/22/1999	60/166,750
2750-0652P	00140.001	11/22/1999	60/166,733
2750-0655P	80173.002	11/23/1999	60/167,362
2750-0654P	80173.001	11/24/1999	60/167,382
2750-0657P	80174.001	11/24/1999	60/167,234
2750-0656P	00141.001	11/24/1999	60/167,233
2750-0658P	80174.002	11/24/1999	60/167,235
2750-0660P	80175.001	11/30/1999	60/167,908
2750-0659P	00142.001	11/30/1999	60/167,904
2750-0661P	80175.002	11/30/1999	60/167,902
2750-0664P	80176.001	12/1/1999	60/168,233
2750-0662P	80042.002	12/1/1999	09/451,320
2750-0665P	80176.002	12/1/1999	60/168,231
2750-0663P	00143.001	12/1/1999	60/168,232
2750-0668P	80177.002	12/2/1999	60/168,548
2750-0667P	80177.001	12/2/1999	60/168,549
2750-0666P	00144.001	12/2/1999	60/168,546
2750-0669P	00145.001	12/3/1999	60/168,675
2750-0670P	80178.001	12/3/1999	60/168,673
2750-0671P	80178.002	12/3/1999	60/168,674
2750-0673P	80179.001	12/7/1999	60/169,278
2750-0672P	00147.001	12/7/1999	60/169,298
2750-0674P	80179.002	12/7/1999	60/169,302
2750-0675P	80180.001	12/8/1999	60/169,692
2750-0676P	80180.002	12/8/1999	60/169,691
2750-0677P	00149.001	12/16/1999	60/171,107
2750-0678P	80181.001	12/16/1999	60/171,114
2750-0679P	80181.002	12/16/1999	60/171,098
2750-0683P	80060.002	1/4/2000	09/478,081
2750-0686F(PC)	80070.100	1/7/2000	00/004,66
2750-0684P	80070.002	1/7/2000	09/479,221
2750-0685P	80183.002	1/19/2000	60/176,867
2750-0688P	80184.002	1/19/2000	60/176,910
2750-0681P	80182.002	1/19/2000	60/176,866
2750-0689P	00152.001	1/26/2000	60/178,166
2750-0691P	80185.001	1/27/2000	60/177,666
2750-0687P	80184.001	1/27/2000	60/178,545
2750-0682P	80183.001	1/27/2000	60/178,546
2750-0680P	80182.001	1/27/2000	60/178,544
2750-0690P	00153.001	1/27/2000	60/178,547
2750-0692P	00155.001	1/28/2000	60/178,754
2750-0693P	80186.001	1/28/2000	60/178,755
2750-0695P	00157.001	2/1/2000	60/179,395
2750-0696P	80187.001	2/1/2000	60/179,388
2750-0694P	80084.002	2/3/2000	09/497,191
2750-0697P	00158.001	2/3/2000	60/180,039

Attorney Docket	Client Docket	Filing Date	Application
2750-0698P	80188.001	2/3/2000	60/180,139
2750-0699P	00159.001	2/4/2000	60/180,206
2750-0700P	80189.001	2/4/2000	60/180,207
2750-0701P	00160.001	2/7/2000	60/180,695
2750-0702P	80190.001	2/7/2000	60/180,696
2750-0703P	00161.001	2/9/2000	60/181,228
2750-0704P	80191.001	2/9/2000	60/181,214
2750-0705P	00162.001	2/10/2000	60/181,476
2750-0706P	80192.001	2/10/2000	60/181,551
2750-0707P	00163.001	2/15/2000	60/182,477
2750-0708P	80193.001	2/15/2000	60/182,516
2750-0712P	00164.001	2/15/2000	60/182,512
2750-0713P	80194.001	2/15/2000	60/182,478
2750-0715P	80195.001	2/17/2000	60/183,165
2750-0714P	00165.001	2/17/2000	60/183,166
2750-0717P	80196.001	2/24/2000	60/184,658
2750-0716P	00167.001	2/24/2000	60/184,667
2750-0709F(CA)	80090.102	2/25/2000	23/006,92
2750-0719P	00168.001	2/25/2000	60/185,118
2750-0718P	91022.001	2/25/2000	60/185,140
2750-0720P	80197.001	2/25/2000	60/185,119
2750-0709F(MX)	80090.101	2/25/2000	00/001,973
2750-0709F(EP)	80090.103	2/25/2000	00/301,439
2750-0709P	80090.002	2/25/2000	09/513,996
2750-0721P	91023.001	2/28/2000	60/185,398
2750-0722P	00169.001	2/28/2000	60/185,396
2750-0723P	80198.001	2/28/2000	60/185,397
2750-0724P	91024.001	2/29/2000	60/185,750
2750-0727P	91025.001	3/1/2000	60/186,277
2750-0725P	00170.001	3/1/2000	
2750-0726P	80199.001	3/1/2000	60/186,296
2750-0710P	80100.002	3/1/2000	09/517,537
2750-0728P	80200.001	3/2/2000	60/187,178
2750-0729P	00172.001	3/2/2000	60/186,386
2750-0730P	80201.001	3/2/2000	60/186,387
2750-0711P	00171.001	3/2/2000	60/186,390
2750-0733P	80202.001	3/3/2000	60/186,669
2750-0731P	91026.001	3/3/2000	60/186,670
2750-0732P	00173.001	3/3/2000	60/186,748
2750-0734P	00174.001	3/7/2000	60/187,378
2750-0735P	91027.001	3/7/2000	60/187,379
2750-0736P	00175.001	3/8/2000	60/187,896
2750-0737P	80203.001	3/8/2000	60/187,888
2750-0738P	91028.001	3/9/2000	60/187,985
2750-0739P	00177.001	3/10/2000	60/188,187
2750-0741P	91030.001	3/10/2000	
2750-0740P	80204.001	3/10/2000	60/188,186
2750-0742P	00178.001	3/10/2000	60/188,185
2750-0743P	80205.001	3/10/2000	60/188,175

Attorney Docket	Client Docket	Filing Date	Application
2750-0744P	91031.001	3/13/2000	60/188,687
2750-0745P	00179.001	3/14/2000	60/189,080
2750-0746P	80206.001	3/14/2000	60/189,052
2750-0749P	80207.001	3/15/2000	60/189,462
2750-0748P	00180.001	3/15/2000	60/189,461
2750-0747P	91032.001	3/15/2000	60/189,460
2750-0753P	80211.001	3/16/2000	60/190,121
2750-0751P	80209.001	3/16/2000	60/189,947
2750-0750P	80208.001	3/16/2000	60/190,120
2750-0756P	80212.001	3/16/2000	60/189,959
2750-0752P	80210.001	3/16/2000	60/189,948
2750-0757P	91034.001	3/16/2000	60/189,965
2750-0754P	91033.001	3/16/2000	60/189,958
2750-0755P	00181.001	3/16/2000	60/189,953
2750-0762P	80214.001	3/20/2000	60/190,089
2750-0761P	00183.001	3/20/2000	60/190,545
2750-0760P	91035.001	3/20/2000	60/190,060
2750-0759P	80213.001	3/20/2000	60/190,070
2750-0758P	00182.001	3/20/2000	60/190,069
2750-0764P	80215.001	3/22/2000	60/191,097
2750-0763P	00184.001	3/22/2000	60/191,084
2750-0766P	00185.001	3/23/2000	60/191,543
2750-0765P	91036.001	3/23/2000	60/191,549
2750-0767P	80216.001	3/23/2000	60/191,545
2750-0770P	80217.001	3/24/2000	60/191,825
2750-0768P	91037.001	3/24/2000	60/191,826
2750-0769P	00186.001	3/24/2000	60/191,823
2750-0772P	00187.001	3/27/2000	60/192,421
2750-0773P	80218.001	3/27/2000	60/192,308
2750-0771P	91038.001	3/27/2000	60/192,420
2750-0774P	91039.001	3/29/2000	60/192,855
2750-0775P	00188.001	3/29/2000	60/192,940
2750-0776P	80219.001	3/29/2000	60/192,941
2750-0778P	00189.001	3/30/2000	60/193,244
2750-0777P	91040.001	3/30/2000	60/193,243
2750-0779P	80220.001	3/30/2000	
2750-0781P	00190.001	3/31/2000	60/193,453
2750-0780P	91041.001	3/31/2000	60/193,469
2750-0782P	80221.001	3/31/2000	60/193,455
2750-0786P	00191.001	4/4/2000	
2750-0787P	80222.001	4/4/2000	
2750-0785P	91042.001	4/4/2000	
2750-0789P	91043.001	4/5/2000	
2750-0790P	00192.001	4/5/2000	
2750-0791P	80223.001	4/5/2000	
2750-0792P	91044.001	4/5/2000	
2750-0783F(CA)	91000.102	4/6/2000	
2750-0783P	91000.002	4/6/2000	
2750-0796P	80225.001	4/6/2000	

Attorney Docket	Client Docket	Filing Date	Application
2750-0783F(EP)	91000.101	4/6/2000	00/302,919
2750-0793P	00193.001	4/6/2000	
2750-0784P	91045.001	4/6/2000	
2750-0795P	00194.001	4/6/2000	
2750-0794P	80224.001	4/6/2000	
2750-0783F(MX)	91000.100	4/6/2000	00/003,391
2750-0799P	80226.001	4/7/2000	
2750-0797P	91046.001	4/7/2000	
2750-0798P	00195.001	4/7/2000	60/195,283
2750-0804P	80228.001	4/11/2000	
2750-0801P	80227.002	4/11/2000	60/196,211
2750-0802P	91047.001	4/11/2000	60/196,168
2750-0803P	00196.001	4/11/2000	
2750-0800P	80227.001	4/12/2000	60/196,212
2750-0805P	91048.001	4/12/2000	60/196,483
2750-0806P	00197.001	4/12/2000	60/196,487
2750-0807P	80229.001	4/12/2000	
2750-0809P	80230.001	4/12/2000	60/196,486
2750-0808P	00200.001	4/12/2000	60/196,485
2750-0811P	80231.002	4/13/2000	
2750-0814P	91049.001	4/14/2000	60/197,397
2750-0810P	80231.001	4/14/2000	
2750-0813P	80232.002	4/17/2000	60/197,871
2750-0812P	80232.001	4/17/2000	60/197,870
2750-0817P	91050.001	4/17/2000	60/198,268
2750-0816P	80233.001	4/17/2000	60/197,678
2750-0815P	00201.001	4/17/2000	60/197,687
2750-0819P	80234.001	4/17/2000	60/197,671
2750-0818P	00202.001	4/17/2000	60/198,133
2750-0820P	91051.001	4/19/2000	60/198,400
2750-0821P	00203.001	4/19/2000	60/198,386
2750-0822P	80235.001	4/19/2000	60/198,373
2750-0823P	91052.001	4/20/2000	60/198,629
2750-0824P	00204.001	4/20/2000	60/198,619
2750-0825P	80236.001	4/20/2000	60/198,623
2750-0828P	80237.001	4/21/2000	60/198,763
2750-0826P	91053.001	4/21/2000	
2750-0827P	00206.001	4/21/2000	60/198,767
2750-0829P	91054.001	4/24/2000	
2750-0830P	00207.001	4/24/2000	
2750-0831P	80238.001	4/24/2000	
2750-0833P	92002.001	4/26/2000	
2750-0834P	00208.001	4/26/2000	
2750-0835P	80239.001	4/26/2000	
2750-0832P	92001.001	4/26/2000	60/200,034
2750-0837P	80240.001	4/27/2000	
2750-0836P	00210.001	4/27/2000	
2750-0844P	80242.002	4/28/2000	
2750-0846P	80243.002	4/28/2000	

Attorney Docket	Client Docket	Filing Date	Application
2750-0788P	80123.002	4/28/2000	
2750-0848P	80244.002	5/1/2000	
2750-0839P	80241.001	5/1/2000	
2750-0845P	80243.001	5/1/2000	
2750-0847P	80244.001	5/1/2000	
2750-0840P	91055.001	5/1/2000	
2750-0843P	80242.001	5/1/2000	
2750-0842P	92002.002	5/1/2000	
2750-0841P	92001.002	5/1/2000	
2750-0838P	00211.001	5/2/2000	
2750-0850P	80245.001	5/2/2000	
2750-0849P	91056.001	5/2/2000	
2750-0852P	80126.002	5/4/2000	
2750-0858P	80246.001	5/4/2000	
2750-0857P	00212.001	5/4/2000	
2750-0856P	91057.001	5/4/2000	
2750-0855P	80130.002	5/5/2000	
2750-0861P	80247.001	5/5/2000	
2750-0859P	91058.001	5/5/2000	
2750-0851F(MX)	91002.102	5/5/2000	
2750-0860P	00213.001	5/5/2000	
2750-0851F(EP)	91002.101	5/5/2000	
2750-0853P	80127.002	5/5/2000	
2750-0854P	80129.002	5/5/2000	
2750-0851F(CA)	91002.100	5/5/2000	
2750-0851P	91002.002	5/5/2000	
2750-0865P	00215.001	5/9/2000	
2750-0866P	80249.001	5/9/2000	
2750-0862P	00214.001	5/9/2000	
2750-0863P	80248.001	5/9/2000	
2750-0864P	91059.001	5/9/2000	
2750-0879P	80252.001	5/10/2000	
2750-0877P	91060.001	5/10/2000	
2750-0878P	00216.001	5/10/2000	
2750-0869P	80251.001	5/11/2000	
2750-0870P	80251.002	5/11/2000	
2750-0867P	80250.001	5/11/2000	
2750-0868P	80250.002	5/11/2000	
2750-0871P	80131.002	5/11/2000	
2750-0880P	91061.001	5/11/2000	
2750-0882P	80253.001	5/11/2000	
2750-0881P	00217.001	5/11/2000	
2750-0875P	91006.002	5/12/2000	
2750-0875F(CA)	91006.100	5/12/2000	
2750-0875F(EP)	91006.101	5/12/2000	
2750-0875F(MX)	91006.102	5/12/2000	
2750-0874P	80116.002	5/12/2000	
2750-0872P	80117.002	5/12/2000	
2750-0883P	91062.001	5/12/2000	

Attorney Docket	Client Docket	Filing Date	Application
2750-0885P	80254.001	5/12/2000	
2750-0884P	00219.001	5/12/2000	
2750-0873P	00025.002	5/12/2000	
2750-0887P	00220.001	5/15/2000	
2750-0888P	80255.001	5/15/2000	
2750-0886P	91063.001	5/15/2000	
2750-0891P	00221.001	5/16/2000	
2750-0892P	80256.001	5/16/2000	
2750-0889P	92001.003	5/17/2000	
2750-0893P	00222.001	5/17/2000	
2750-0894P	80257.001	5/17/2000	
2750-0890P	92002.003	5/17/2000	
2750-0895P	00223.001	5/18/2000	
2750-0876F(MX)	91007.102	5/18/2000	
2750-0876F(EP)	91007.101	5/18/2000	
2750-0876F(CA)	91007.100	5/18/2000	
2750-0896P	80258.001	5/18/2000	
2750-0876P	91007.002	5/18/2000	
2750-0897P	00224.001	5/19/2000	
2750-0898P	80259.001	5/19/2000	
2750-0901P	80260.001	5/22/2000	
2750-0900P	00225.001	5/22/2000	
2750-0899P	91064.001	5/22/2000	
2750-0903P	80261.001	5/23/2000	
2750-0902P	00226.001	5/23/2000	
2750-0904P	00227.001	5/24/2000	
2750-0905P	80262.001	5/24/2000	
2750-0906P	91065.001	5/25/2000	
2750-0907P	00228.001	5/26/2000	
2750-0911P	80264.001	5/26/2000	
2750-0910P	00229.001	5/26/2000	
2750-0908P	80263.001	5/26/2000	
2750-0909P	91066.001	5/26/2000	
2750-0913P	00230.001	5/30/2000	
2750-0914P	80265.001	5/30/2000	
2750-0912P	91067.001	5/30/2000	
2750-0921P	80268.001	6/1/2000	
2750-0920P	00231.001	6/1/2000	
2750-0919P	91068.001	6/1/2000	
2750-0918P	80267.002	6/1/2000	
2750-0916P	80266.002	6/1/2000	
2750-0915P	80266.001	6/2/2000	
2750-0917P	80267.001	6/2/2000	
2750-0922P	91069.001	6/5/2000	
2750-0923P	00232.001	6/5/2000	
2750-0924P	80269.001	6/5/2000	
2750-0925P	91070.001	6/5/2000	
2750-0926P	00233.001	6/5/2000	
2750-0927P	80270.001	6/5/2000	

Attorney Docket	Client Docket	Filing Date	Application
2750-0928P	00033.003	6/9/2000	
2750-0929P	91071.001	6/8/2000	
2750-0930P	00234.001	6/8/2000	
2750-0931P	80271.001	6/8/2000	
2750-0932P	00235.001	6/9/2000	
2750-0933P	80272.001	6/9/2000	

All applications listed in the table above are expressly incorporated herein by reference in their entirety and for all purposes.

The SDFs of the invention can also be used as probes to search for genes that are related to the SDF within a species. Such related genes are typically considered to be members of a gene family. In such a case, the sequence similarity will often be concentrated into one or a few fragments of the sequence. The fragments of similar sequence that define the gene family typically encode a fragment of a protein or RNA that has an enzymatic or structural function. The percentage of identity in the amino acid sequence of the domain that defines the gene family is preferably at least 70%, more preferably 80 to 95%, most preferably 85 to 99%. To search for members of a gene family within a species, a low stringency hybridization is usually performed, but this will depend upon the size, distribution and degree of sequence divergence of domains that define the gene family. SDFs encompassing regulatory regions can be used to identify coordinately expressed genes by using the regulatory region sequence of the SDF as a probe.

In the instances where the SDFs are identified as being expressed from genes that confer a particular phenotype, then the SDFs can also be used as probes to assay plants of different species for those phenotypes.

I.C. Methods to Inhibit Gene Expression

The nucleic acid molecules of the present invention can be used to inhibit gene transcription and/or translation. Example of such methods include, without limitation:

Antisense Constructs;

Ribozyme Constructs;

Chimeraplast Constructs;

Co-Suppression;

Transcriptional Silencing; and

Other Methods of Gene Expression.

C.1 Antisense

In some instances it is desirable to suppress expression of an endogenous or
5 exogenous gene. A well-known instance is the FLAVOR-SAVOR™ tomato, in which the
gene encoding ACC synthase is inactivated by an antisense approach, thus delaying softening
of the fruit after ripening. See for example, U.S. Patent No. 5,859,330; U.S. Patent No.
5,723,766; Oeller, et al, *Science*, 254:437-439(1991); and Hamilton et al, *Nature*, 346:284-
287 (1990). Also, timing of flowering can be controlled by suppression of the *FLOWERING*
10 *LOCUS C (FLC)*; high levels of this transcript are associated with late flowering, while
absence of *FLC* is associated with early flowering (S.D. Michaels et al., *Plant Cell* 11:949
(1999). Also, the transition of apical meristem from production of leaves with associated
shoots to flowering is regulated by *TERMINAL FLOWER1*, *APETALA1* and *LEAFY*. Thus,
when it is desired to induce a transition from shoot production to flowering, it is desirable to
15 suppress *TFL1* expression (S.J. Liljegren, *Plant Cell* 11:1007 (1999)). As another instance,
arrested ovule development and female sterility result from suppression of the ethylene
forming enzyme but can be reversed by application of ethylene (D. De Martinis et al., *Plant*
Cell 11:1061 (1999)). The ability to manipulate female fertility of plants is useful in
increasing fruit production and creating hybrids.

20 In the case of polynucleotides used to inhibit expression of an endogenous gene, the
introduced sequence need not be perfectly identical to a sequence of the target endogenous gene.
The introduced polynucleotide sequence will typically be at least substantially identical to the
target endogenous sequence.

Some polynucleotide SDFs in Table 1 represent sequences that are expressed in
25 corn, wheat, rice, soybean *Arabidopsis* and/or other plants. Thus the invention includes using
these sequences to generate antisense constructs to inhibit translation and/or degradation of
transcripts of said SDFs, typically in a plant cell.

To accomplish this, a polynucleotide segment from the desired gene that can hybridize to
the mRNA expressed from the desired gene (the “antisense segment”) is operably linked to a
30 promoter such that the antisense strand of RNA will be transcribed when the construct is present
in a host cell. A regulated promoter can be used in the construct to control transcription of the
antisense segment so that transcription occurs only under desired circumstances.

The antisense segment to be introduced generally will be substantially identical to at least a fragment of the endogenous gene or genes to be repressed. The sequence, however, need not be perfectly identical to inhibit expression. Further, the antisense product may hybridize to the untranslated region instead of or in addition to the coding sequence of the gene. The vectors of the present invention can be designed such that the inhibitory effect applies to other proteins within a family of genes exhibiting homology or substantial homology to the target gene.

For antisense suppression, the introduced antisense segment sequence also need not be full length relative to either the primary transcription product or the fully processed mRNA. Generally, a higher percentage of sequence identity can be used to compensate for the use of a shorter sequence. Furthermore, the introduced sequence need not have the same intron or exon pattern, and homology of non-coding segments may be equally effective. Normally, a sequence of between about 30 or 40 nucleotides and the full length of the transcript can be used, though a sequence of at least about 100 nucleotides is preferred, a sequence of at least about 200 nucleotides is more preferred, and a sequence of at least about 500 nucleotides is especially preferred.

C.2. Ribozymes

It is also contemplated that gene constructs representing ribozymes and based on the SDFs in TABLE 1 are an object of the invention. Ribozymes can also be used to inhibit expression of genes by suppressing the translation of the mRNA into a polypeptide. It is possible to design ribozymes that specifically pair with virtually any target RNA and cleave the phosphodiester backbone at a specific location, thereby functionally inactivating the target RNA. In carrying out this cleavage, the ribozyme is not itself altered, and is thus capable of recycling and cleaving other molecules, making it a true enzyme. The inclusion of ribozyme sequences within antisense RNAs confers RNA-cleaving activity upon them, thereby increasing the activity of the constructs.

A number of classes of ribozymes have been identified. One class of ribozymes is derived from a number of small circular RNAs, which are capable of self-cleavage and replication in plants. The RNAs replicate either alone (viroid RNAs) or with a helper virus (satellite RNAs). Examples include RNAs from avocado sunblotch viroid and the satellite RNAs from tobacco ringspot virus, lucerne transient streak virus, velvet tobacco mottle virus,

solanum nodiflorum mottle virus and subterranean clover mottle virus. The design and use of target RNA-specific ribozymes is described in Haseloff et al. *Nature*, 334:585 (1988).

Like the antisense constructs above, the ribozyme sequence fragment necessary for pairing need not be identical to the target nucleotides to be cleaved, nor identical to the sequences in TABLE 1. Ribozymes may be constructed by combining the ribozyme sequence and some fragment of the target gene which would allow recognition of the target gene mRNA by the resulting ribozyme molecule. Generally, the sequence in the ribozyme capable of binding to the target sequence exhibits a percentage of sequence identity with at least 80%, preferably with at least 85%, more preferably with at least 90% and most preferably with at least 95%, even more preferably, with at least 96%, 97%, 98% or 99% sequence identity to some fragment of a sequence in TABLE 1 or the complement thereof. The ribozyme can be equally effective in inhibiting mRNA translation by cleaving either in the untranslated or coding regions. Generally, a higher percentage of sequence identity can be used to compensate for the use of a shorter sequence. Furthermore, the introduced sequence need not have the same intron or exon pattern, and homology of non-coding segments may be equally effective.

C.3. Chimeraplasts

The SDFs of the invention, such as those described by Table 1, can also be used to construct chimeraplasts that can be introduced into a cell to produce at least one specific nucleotide change in a sequence corresponding to the SDF of the invention. A chimeraplast is an oligonucleotide comprising DNA and/or RNA that specifically hybridizes to a target region in a manner which creates a mismatched base-pair. This mismatched base-pair signals the cell's repair enzyme machinery which acts on the mismatched region resulting in the replacement, insertion or deletion of designated nucleotide(s). The altered sequence is then expressed by the cell's normal cellular mechanisms. Chimeraplasts can be designed to repair mutant genes, modify genes, introduce site-specific mutations, and/or act to interrupt or alter normal gene function (US Pat. Nos. 6,010,907 and 6,004,804; and PCT Pub. No. WO99/58723 and WO99/07865).

C.4. Sense Suppression

The SDFs of Table 1 of the present invention are also useful to modulate gene expression by sense suppression. Sense suppression represents another method of gene

suppression by introducing at least one exogenous copy or fragment of the endogenous sequence to be suppressed.

Introduction of expression cassettes in which a nucleic acid is configured in the sense orientation with respect to the promoter into the chromosome of a plant or by a self-replicating virus has been shown to be an effective means by which to induce degradation of mRNAs of target genes. For an example of the use of this method to modulate expression of endogenous genes see, Napoli et al., *The Plant Cell* 2:279 (1990), and U.S. Patents Nos. 5,034,323, 5,231,020, and 5,283,184. Inhibition of expression may require some transcription of the introduced sequence.

For sense suppression, the introduced sequence generally will be substantially identical to the endogenous sequence intended to be inactivated. The minimal percentage of sequence identity will typically be greater than about 65%, but a higher percentage of sequence identity might exert a more effective reduction in the level of normal gene products. Sequence identity of more than about 80% is preferred, though about 95% to absolute identity would be most preferred. As with antisense regulation, the effect would likely apply to any other proteins within a similar family of genes exhibiting homology or substantial homology to the suppressing sequence.

C.5. Transcriptional Silencing

The nucleic acid sequences of the invention, including the SDFs of Table 1, and fragments thereof, contain sequences that can be inserted into the genome of an organism resulting in transcriptional silencing. Such regulatory sequences need not be operatively linked to coding sequences to modulate transcription of a gene. Specifically, a promoter sequence without any other element of a gene can be introduced into a genome to transcriptionally silence an endogenous gene (see, for example, Vaucheret, H et al. (1998) *The Plant Journal* 16: 651-659). As another example, triple helices can be formed using oligonucleotides based on sequences from TABLE 1, fragments thereof, and substantially similar sequence thereto. The oligonucleotide can be delivered to the host cell and can bind to the promoter in the genome to form a triple helix and prevent transcription. An oligonucleotide of interest is one that can bind to the promoter and block binding of a transcription factor to the promoter. In such a case, the oligonucleotide can be complementary to the sequences of the promoter that interact with transcription binding factors.

C.6. Other Methods to Inhibit Gene Expression

Yet another means of suppressing gene expression is to insert a polynucleotide into the gene of interest to disrupt transcription or translation of the gene.

5 Low frequency homologous recombination can be used to target a polynucleotide insert to a gene by flanking the polynucleotide insert with sequences that are substantially similar to the gene to be disrupted. Sequences from TABLE 1, fragments thereof, and substantially similar sequence thereto can be used for homologous recombination.

10 In addition, random insertion of polynucleotides into a host cell genome can also be used to disrupt the gene of interest. Azpiroz-Leehan et al., *Trends in Genetics* 13:152 (1997). In this method, screening for clones from a library containing random insertions is preferred to identifying those that have polynucleotides inserted into the gene of interest. Such screening can be performed using probes and/or primers described above based on sequences from TABLE 1, fragments thereof, and substantially similar sequence thereto. The screening can also be
15 performed by selecting clones or R₁ plants having a desired phenotype.

I.D. Methods of Functional Analysis

The constructs described in the methods under I.C. above can be used to determine the function of the polypeptide encoded by the gene that is targeted by the constructs.

20 Down-regulating the transcription and translation of the targeted gene in the host cell or organisms, such as a plant, may produce phenotypic changes as compared to a wild-type cell or organism. In addition, *in vitro* assays can be used to determine if any biological activity, such as calcium flux, DNA transcription, nucleotide incorporation, etc., are being modulated by the down-regulation of the targeted gene.

25 Coordinated regulation of sets of genes, e.g., those contributing to a desired polygenic trait, is sometimes necessary to obtain a desired phenotype. SDFs of the invention representing transcription activation and DNA binding domains can be assembled into hybrid transcriptional activators. These hybrid transcriptional activators can be used with their corresponding DNA elements (i.e., those bound by the DNA-binding SDFs) to effect coordinated expression of desired genes (J.J. Schwarz et al., *Mol. Cell. Biol.* 12:266 (1992),
30 A. Martinez et al., *Mol. Gen. Genet.* 261:546 (1999)).

The SDFs of the invention can also be used in the two-hybrid genetic systems to identify networks of protein-protein interactions (L. McAlister-Henn et al., *Methods* 19:330 (1999), J.C. Hu et al., *Methods* 20:80 (2000), M. Golovkin et al., *J. Biol. Chem.* 274:36428 (1999), K. Ichimura et al., *Biochem. Biophys. Res. Comm.* 253:532 (1998)). The SDFs of the invention can also be used in various expression display methods to identify important protein-DNA interactions (e.g. B. Luo et al., *J. Mol. Biol.* 266:479 (1997)).

I.E. Promoters

The SDFs of the invention are also useful as structural or regulatory sequences in a construct for modulating the expression of the corresponding gene in a plant or other organism, e.g. a symbiotic bacterium. For example, promoter sequences associated to SDFs of Table 1 of the present invention can be useful in directing expression of coding sequences either as constitutive promoters or to direct expression in particular cell types, tissues, or organs or in response to environmental stimuli.

With respect to the SDFs of the present invention a promoter is likely to be a relatively small portion of a genomic DNA (gDNA) sequence located in the first 2000 nucleotides upstream from an initial exon identified in a gDNA sequence or initial "ATG" or methionine codon or translational start site in a corresponding cDNA sequence. Such promoters are more likely to be found in the first 1000 nucleotides upstream of an initial ATG or methionine codon or translational start site of a cDNA sequence corresponding to a gDNA sequence. In particular, the promoter is usually located upstream of the transcription start site. The fragments of a particular gDNA sequence that function as elements of a promoter in a plant cell will preferably be found to hybridize to gDNA sequences presented and described in Table 1 at medium or high stringency, relevant to the length of the probe and its base composition.

Promoters are generally modular in nature. Promoters can consist of a basal promoter that functions as a site for assembly of a transcription complex comprising an RNA polymerase, for example RNA polymerase II. A typical transcription complex will include additional factors such as TF_{II}B, TF_{II}D, and TF_{II}E. Of these, TF_{II}D appears to be the only one to bind DNA directly. The promoter might also contain one or more enhancers and/or suppressors that function as binding sites for additional transcription factors that have the function of modulating

the level of transcription with respect to tissue specificity and of transcriptional responses to particular environmental or nutritional factors, and the like.

Short DNA sequences representing binding sites for proteins can be separated from each other by intervening sequences of varying length. For example, within a particular functional module, protein binding sites may be constituted by regions of 5 to 60, preferably 10 to 30, more preferably 10 to 20 nucleotides. Within such binding sites, there are typically 2 to 6 nucleotides that specifically contact amino acids of the nucleic acid binding protein. The protein binding sites are usually separated from each other by 10 to several hundred nucleotides, typically by 15 to 150 nucleotides, often by 20 to 50 nucleotides. DNA binding sites in promoter elements often display dyad symmetry in their sequence. Often elements binding several different proteins, and/or a plurality of sites that bind the same protein, will be combined in a region of 50 to 1,000 basepairs.

Elements that have transcription regulatory function can be isolated from their corresponding endogenous gene, or the desired sequence can be synthesized, and recombined in constructs to direct expression of a coding region of a gene in a desired tissue-specific, temporal-specific or other desired manner of inducibility or suppression. When hybridizations are performed to identify or isolate elements of a promoter by hybridization to the long sequences presented in TABLE 1, conditions are adjusted to account for the above-described nature of promoters. For example short probes, constituting the element sought, are preferably used under low temperature and/or high salt conditions. When long probes, which might include several promoter elements are used, low to medium stringency conditions are preferred when hybridizing to promoters across species.

If a nucleotide sequence of an SDF, or part of the SDF, functions as a promoter or fragment of a promoter, then nucleotide substitutions, insertions or deletions that do not substantially affect the binding of relevant DNA binding proteins would be considered equivalent to the exemplified nucleotide sequence. It is envisioned that there are instances where it is desirable to decrease the binding of relevant DNA binding proteins to silence or down-regulate a promoter, or conversely to increase the binding of relevant DNA binding proteins to enhance or up-regulate a promoter and vice versa. In such instances, polynucleotides representing changes to the nucleotide sequence of the DNA-protein contact region by insertion of additional nucleotides, changes to identity of relevant nucleotides, including use of chemically-modified bases, or deletion of one or more nucleotides are considered encompassed by the present invention. In addition, fragments of the promoter

sequences described by Table 1 and variants thereof can be fused with other promoters or fragments to facilitate transcription and/or transcription in specific type of cells or under specific conditions.

Promoter function can be assayed by methods known in the art, preferably by measuring activity of a reporter gene operatively linked to the sequence being tested for promoter function. Examples of reporter genes include those encoding luciferase, green fluorescent protein, GUS, neo, cat and bar.

I.F. UTRs and Junctions

Polynucleotides comprising untranslated (UTR) sequences and intron/exon junctions are also within the scope of the invention. UTR sequences include introns and 5' or 3' untranslated regions (5' UTRs or 3' UTRs). Fragments of the sequences shown in TABLE 1 can comprise UTRs and intron/exon junctions.

These fragments of SDFs, especially UTRs, can have regulatory functions related to, for example, translation rate and mRNA stability. Thus, these fragments of SDFs can be isolated for use as elements of gene constructs for regulated production of polynucleotides encoding desired polypeptides.

Introns of genomic DNA segments might also have regulatory functions. Sometimes regulatory elements, especially transcription enhancer or suppressor elements, are found within introns. Also, elements related to stability of heteronuclear RNA and efficiency of splicing and of transport to the cytoplasm for translation can be found in intron elements. Thus, these segments can also find use as elements of expression vectors intended for use to transform plants.

Just as with promoters UTR sequences and intron/exon junctions can vary from those shown in TABLE 1. Such changes from those sequences preferably will not affect the regulatory activity of the UTRs or intron/exon junction sequences on expression, transcription, or translation unless selected to do so. However, in some instances, down- or up-regulation of such activity may be desired to modulate traits or phenotypic or *in vitro* activity.

I.G. Coding Sequences

Isolated polynucleotides of the invention can include coding sequences that encode polypeptides comprising an amino acid sequence encoded by sequences in TABLE 1 or an amino acid sequence presented in TABLE 1.

A nucleotide sequence encodes a polypeptide if a cell (or a cell free *in vitro* system) expressing that nucleotide sequence produces a polypeptide having the recited amino acid sequence when the nucleotide sequence is transcribed and the primary transcript is subsequently processed and translated by a host cell (or a cell free *in vitro* system) harboring the nucleic acid. Thus, an isolated nucleic acid that encodes a particular amino acid sequence can be a genomic sequence comprising exons and introns or a cDNA sequence that represents the product of splicing thereof. An isolated nucleic acid encoding an amino acid sequence also encompasses heteronuclear RNA, which contains sequences that are spliced out during expression, and mRNA, which lacks those sequences.

Coding sequences can be constructed using chemical synthesis techniques or by isolating coding sequences or by modifying such synthesized or isolated coding sequences as described above.

In addition to coding sequences encoding the polypeptide sequences of TABLE 1, which are native to corn, *Arabidopsis*, soybean, rice, wheat, and other plants the isolated polynucleotides can be polynucleotides that encode variants, fragments, and fusions of those native proteins. Such polypeptides are described below in part II.

In variant polynucleotides generally, the number of substitutions, deletions or insertions is preferably less than 20%, more preferably less than 15%; even more preferably less than 10%, 5%, 3% or 1% of the number of nucleotides comprising a particularly exemplified sequence. It is generally expected that non-degenerate nucleotide sequence changes that result in 1 to 10, more preferably 1 to 5 and most preferably 1 to 3 amino acid insertions, deletions or substitutions will not greatly affect the function of an encoded polypeptide. The most preferred embodiments are those wherein 1 to 20, preferably 1 to 10, most preferably 1 to 5 nucleotides are added to, deleted from and/or substituted in the sequences specifically disclosed in TABLE 1.

Insertions or deletions in polynucleotides intended to be used for encoding a polypeptide preferably preserve the reading frame. This consideration is not so important in instances when the polynucleotide is intended to be used as a hybridization probe.

II. Polypeptides and Proteins

IIA. Native polypeptides and proteins

Polypeptides within the scope of the invention include both native proteins as well as variants, fragments, and fusions thereof. Polypeptides of the invention are those encoded by any of the six reading frames of sequences shown in TABLE 1, preferably encoded by the three frames reading in the 5' to 3' direction of the sequences as shown.

Native polypeptides include the proteins encoded by the sequences shown in TABLE 1. Such native polypeptides include those encoded by allelic variants.

Polypeptide and protein variants will exhibit at least 75% sequence identity to those native polypeptides of TABLE 1. More preferably, the polypeptide variants will exhibit at least 85% sequence identity; even more preferably, at least 90% sequence identity; more preferably at least 95%, 96%, 97%, 98%, or 99% sequence identity. Fragments of polypeptide or fragments of polypeptides will exhibit similar percentages of sequence identity to the relevant fragments of the native polypeptide. Fusions will exhibit a similar percentage of sequence identity in that fragment of the fusion represented by the variant of the native peptide.

Furthermore, polypeptide variants will exhibit at least one of the functional properties of the native protein. Such properties include, without limitation, protein interaction, DNA interaction, biological activity, immunological activity, receptor binding, signal transduction, transcription activity, growth factor activity, secondary structure, three-dimensional structure, etc. As to properties related to *in vitro* or *in vivo* activities, the variants preferably exhibit at least 60% of the activity of the native protein; more preferably at least 70%, even more preferably at least 80%, 85%, 90% or 95% of at least one activity of the native protein.

One type of variant of native polypeptides comprises amino acid substitutions, deletions and/or insertions. Conservative substitutions are preferred to maintain the function or activity of the polypeptide.

Within the scope of percentage of sequence identity described above, a polypeptide of the invention may have additional individual amino acids or amino acid sequences inserted into the polypeptide in the middle thereof and/or at the N-terminal and/or C-terminal ends thereof. Likewise, some of the amino acids or amino acid sequences may be deleted from the polypeptide.

A.1 Antibodies

Isolated polypeptides can be utilized to produce antibodies. Polypeptides of the invention can generally be used, for example, as antigens for raising antibodies by known techniques. The resulting antibodies are useful as reagents for determining the distribution of the antigen protein within the tissues of a plant or within a cell of a plant. The antibodies are also useful for examining the production level of proteins in various tissues, for example in a wild-type plant or following genetic manipulation of a plant, by methods such as Western blotting.

Antibodies of the present invention, both polyclonal and monoclonal, may be prepared by conventional methods. In general, the polypeptides of the invention are first used to immunize a suitable animal, such as a mouse, rat, rabbit, or goat. Rabbits and goats are preferred for the preparation of polyclonal sera due to the volume of serum obtainable, and the availability of labeled anti-rabbit and anti-goat antibodies as detection reagents. Immunization is generally performed by mixing or emulsifying the protein in saline, preferably in an adjuvant such as Freund's complete adjuvant, and injecting the mixture or emulsion parenterally (generally subcutaneously or intramuscularly). A dose of 50-200 µg/injection is typically sufficient. Immunization is generally boosted 2-6 weeks later with one or more injections of the protein in saline, preferably using Freund's incomplete adjuvant. One may alternatively generate antibodies by *in vitro* immunization using methods known in the art, which for the purposes of this invention is considered equivalent to *in vivo* immunization.

Polyclonal antisera is obtained by bleeding the immunized animal into a glass or plastic container, incubating the blood at 25°C for one hour, followed by incubating the blood at 4°C for 2-18 hours. The serum is recovered by centrifugation (e.g., 1,000xg for 10 minutes). About 20-50 ml per bleed may be obtained from rabbits.

Monoclonal antibodies are prepared using the method of Kohler and Milstein, *Nature* 256: 495 (1975), or modification thereof. Typically, a mouse or rat is immunized as described above. However, rather than bleeding the animal to extract serum, the spleen (and optionally several large lymph nodes) is removed and dissociated into single cells. If desired, the spleen cells can be screened (after removal of nonspecifically adherent cells) by applying a cell suspension to a plate, or well, coated with the protein antigen. B-cells producing membrane-bound immunoglobulin specific for the antigen bind to the plate, and are not rinsed away with the rest of the suspension. Resulting B-cells, or all dissociated spleen cells, are then induced to fuse with myeloma cells to form hybridomas, and are cultured in a selective medium (e.g., hypoxanthine, aminopterin, thymidine medium, "HAT"). The resulting hybridomas are plated

by limiting dilution, and are assayed for the production of antibodies which bind specifically to the immunizing antigen (and which do not bind to unrelated antigens). The selected Mab-secreting hybridomas are then cultured either *in vitro* (e.g., in tissue culture bottles or hollow fiber reactors), or *in vivo* (as ascites in mice).

5 Other methods for sustaining antibody-producing B-cell clones, such as by EBV transformation, are known.

If desired, the antibodies (whether polyclonal or monoclonal) may be labeled using conventional techniques. Suitable labels include fluorophores, chromophores, radioactive atoms (particularly ^{32}P and ^{125}I), electron-dense reagents, enzymes, and ligands having specific binding
10 partners. Enzymes are typically detected by their activity. For example, horseradish peroxidase is usually detected by its ability to convert 3,3',5,5'-tetramethylbenzidine (TNB) to a blue pigment, quantifiable with a spectrophotometer.

A.2 In Vitro Applications of Polypeptides

Some polypeptides of the invention will have enzymatic activities that are useful *in vitro*.
15 For example, the soybean trypsin inhibitor (Kunitz) family is one of the numerous families of proteinase inhibitors. It comprises plant proteins which have inhibitory activity against serine proteinases from the trypsin and subtilisin families, thiol proteinases and aspartic proteinases. Thus, these peptides find *in vitro* use in protein purification protocols and perhaps in therapeutic settings requiring topical application of protease inhibitors.

20 Delta-aminolevulinic acid dehydratase (EC 4.2.1.24) (ALAD) catalyzes the second step in the biosynthesis of heme, the condensation of two molecules of 5-aminolevulinate to form porphobilinogen and is also involved in chlorophyll biosynthesis (Kaczor et al. (1994) Plant Physiol. 1-4: 1411-7; Smith (1988) Biochem. J. 249: 423-8; Schneider (1976) Z. naturforsch. [C] 31: 55-63). Thus, ALAD proteins can be used as catalysts in synthesis of
25 heme derivatives. Enzymes of biosynthetic pathways generally can be used as catalysts for *in vitro* synthesis of the compounds representing products of the pathway.

Polypeptides encoded by SDFs of the invention can be engineered to provide purification reagents to identify and purify additional polypeptides that bind to them. This allows one to identify proteins that function as multimers or elucidate signal transduction or
30 metabolic pathways. In the case of DNA binding proteins, the polypeptide can be used in a similar manner to identify the DNA determinants of specific binding (S. Pierrou et al., *Anal.*

Biochem. 229:99 (1995), S. Chusacultachai et al., *J. Biol. Chem.* 274:23591 (1999), Q. Lin et al., *J. Biol. Chem.* 272:27274 (1997)).

II.B. POLYPEPTIDE VARIANTS, FRAGMENTS, AND FUSIONS

Generally, variants, fragments, or fusions of the polypeptides encoded by the SDFs of the invention can exhibit at least one of the activities of the identified domains and/or related polypeptides described in Table 1 corresponding to the SDF of interest.

II.B.(1) Variants

A type of variant of the native polypeptides comprises amino acid substitutions. Conservative substitutions, described above (see II.), are preferred to maintain the function or activity of the polypeptide. Such substitutions include conservation of charge, polarity, hydrophobicity, size, etc. For example, one or more amino acid residues within the sequence can be substituted with another amino acid of similar polarity that acts as a functional equivalent, for example providing a hydrogen bond in an enzymatic catalysis. Substitutes for an amino acid within an exemplified sequence are preferably made among the members of the class to which the amino acid belongs. For example, the nonpolar (hydrophobic) amino acids include alanine, leucine, isoleucine, valine, proline, phenylalanine, tryptophan and methionine. The polar neutral amino acids include glycine, serine, threonine, cysteine, tyrosine, asparagine, and glutamine. The positively charged (basic) amino acids include arginine, lysine and histidine. The negatively charged (acidic) amino acids include aspartic acid and glutamic acid.

Within the scope of percentage of sequence identity described above, a polypeptide of the invention may have additional individual amino acids or amino acid sequences inserted into the polypeptide in the middle thereof and/or at the N-terminal and/or C-terminal ends thereof. Likewise, some of the amino acids or amino acid sequences may be deleted from the polypeptide. Amino acid substitutions may also be made in the sequences; conservative substitutions being preferred.

One preferred class of variants are those that comprise (1) the domain of an encoded polypeptide and/or (2) residues conserved between the encoded polypeptide and related polypeptides. For this class of variants, the encoded polypeptide sequence is changed by insertion, deletion, or substitution at positions flanking the domain and/or conserved residues.

Another class of variants includes those that comprise an encoded polypeptide sequence that is changed in the domain or conserved residues by a conservative substitution.

Yet another class of variants includes those that lack one of the *in vitro* activities, or structural features of the encoded polypeptides. One example is polypeptides or proteins produced from genes comprising dominant negative mutations. Such a variant may comprise an encoded polypeptide sequence with non-conservative changes in a particular domain or group of conserved residues.

II.A.(2) FRAGMENTS

Fragments of particular interest are those that comprise a domain identified for a polypeptide encoded by an SDF of the instant invention and variants thereof. Also, fragments that comprise at least one region of residues conserved between an SDF encoded polypeptide and its related polypeptides are of great interest. Fragments are sometimes useful as polypeptides corresponding to genes comprising dominant negative mutations are.

II.A.(3) FUSIONS

Of interest are chimeras comprising (1) a fragment of the SDF encoded polypeptide or variants thereof of interest and (2) a fragment of a polypeptide comprising the same domain. For example, an AP2 helix encoded by a SDF of the invention fused to second AP2 helix from ANT protein, which comprises two AP2 helices. The present invention also encompasses fusions of SDF encoded polypeptides, variants, or fragments thereof fused with related proteins or fragments thereof.

DEFINITION OF DOMAINS

The polypeptides of the invention may possess identifying domains. In addition, the domains within the SDF encoded polypeptide can be defined by the region that exhibits at least 70% sequence identity with the consensus sequences listed in the detailed description below of each of the domains.

The majority of the protein domain descriptions given below are obtained from Prosite, (<http://www.expasy.ch/prosite/>), and Pfam, (<http://pfam.wustl.edu/browse.shtml>).

1. (AAA) AAA-protein family signature

A large family of ATPases has been described [1 to 5] whose key feature is that they share a conserved region of about 220 amino acids that contains an ATP-binding site. This family is now called AAA, for 'A'TPases 'A'ssociated with diverse cellular 'A'ctivities. The proteins that belong to this family either contain one or two AAA domains. Proteins containing two AAA domains:

- Mammalian and drosophila NSF (N-ethylmaleimide-sensitive fusion protein) and the fungal homolog, SEC18. These proteins are involved in intracellular transport between the endoplasmic reticulum and Golgi, as well as between different Golgi cisternae.
- Mammalian transitional endoplasmic reticulum ATPase (previously known as p97 or VCP) which is involved in the transfer of membranes from the endoplasmic reticulum to the golgi apparatus. This protein forms a ring-shaped homooligomer composed of six subunits. The yeast homolog is CDC48 and it may play a role in spindle pole proliferation.
- Yeast protein PAS1, essential for peroxisome assembly and the related protein PAS1 from *Pichia pastoris*.
- Yeast protein AFG2.
- *Sulfolobus acidocaldarius* protein SAV and *Halobacterium salinarium* cdcH which may be part of a transduction pathway connecting light to cell division.

Proteins containing a single AAA domain:

- *Escherichia coli* and other bacteria ftsH (or hflB) protein. FtsH is an ATP-dependent zinc metalloprotease that seems to degrade the heat-shock sigma-32 factor.

It is an integral membrane protein with a large cytoplasmic C-terminal domain that contain both the AAA and the protease domains.

- Yeast protein YME1, a protein important for maintaining the integrity of the mitochondrial compartment. YME1 is also a zinc-dependent protease.
- Yeast protein AFG3 (or YTA10). This protein also seems to contain a AAA domain followed by a zinc-dependent protease domain.

Subunits from the regulatory complex of the 26S proteasome [6] which is involved in the ATP-dependent degradation of ubiquitinated proteins:

- a) Mammalian subunit 4 and homologs in other higher eukaryotes, in yeast (gene YTA5) and fission yeast (gene mts2).

- b) Mammalian subunit 6 (TBP7) and homologs in other higher eukaryotes and in yeast (gene YTA2).
- c) Mammalian subunit 7 (MSS1) and homologs in other higher eukaryotes and in yeast (gene CIM5 or YTA3).
- d) Mammalian subunit 8 (P45) and homologs in other higher eukaryotes and in yeast (SUG1 or CIM3 or TBY1) and fission yeast (gene let1).

Other probable subunits such as human TBP1 which seems to influence HIV gene expression by interacting with the virus tat transactivator protein and yeast YTA1 and YTA6.

- Yeast protein BCS1, a mitochondrial protein essential for the expression of the Rieske iron-sulfur protein.
- Yeast protein MSP1, a protein involved in intramitochondrial sorting of proteins.
- Yeast protein PAS8, and the corresponding proteins PAS5 from *Pichia pastoris* and PAY4 from *Yarrowia lipolytica*.
- Mouse protein SKD1 and its fission yeast homolog (SpAC2G11.06).
- *Caenorhabditis elegans* meiotic spindle formation protein mei-1.
- Yeast protein SAP1.
- Yeast protein YTA7.
- *Mycobacterium leprae* hypothetical protein A2126A.

It is proposed that, in general, the AAA domains in these proteins act as ATP-dependent protein clamps [5]. In addition to the ATP-binding 'A' and 'B' motifs, which are located in the N-terminal half of this domain, there is a highly conserved region located in the central part of the domain which was used to develop a signature pattern.

Consensus pattern: [LIVMT SEQ ID NO:1)]-x-[LIVMT SEQ ID NO:1)]-[LIVMF SEQ ID NO:2)]-x-[GATMC SEQ ID NO:3)]-[ST]-[NS]-x(4)-[LIVM SEQ ID NO:4)]-D-x-A-[LIFA SEQ ID NO:5)]-x-R

[1] Froehlich K.-U., Fries H.W., Ruediger M., Erdmann R., Botstein D., Mecke D. J. Cell Biol. 114:443-453(1991).

[2] Erdmann R., Wiebel F.F., Flessau A., Rytka J., Beyer A., Froehlich K.-U., Kunau W.-H. Cell 64:499-510(1991).

[3] Peters J.-M., Walsh M.J., Franke W.W. EMBO J. 9:1757-1767(1990).

[4] Kunau W.-H., Beyer A., Goette K., Marzioch M., Saidowsky J., Skaletz-Rorowski A., Wiebel F.F. *Biochimie* 75:209-224(1993).

[5] Confalonieri F., Duguet M. *BioEssays* 17:639-650(1995). [6] Hilt W., Wolf D.H. *Trends Biochem. Sci.* 21:96-102(1996).

5

2. ABC Membrane (ABC transporter transmembrane region). This family represents a unit of six transmembrane helices. Many members of the ABC transporter family (ABC_tran) have two such regions. See also descriptions of ABC Tran, below, and ABC2 membrane, above.

10

3. (ABC Tran) ABC transporters family signature. On the basis of sequence similarities a family of related ATP-binding proteins has been characterized [1 to 5]. These proteins are associated with a variety of distinct biological processes in both prokaryotes and eukaryotes, but a majority of them are involved in active transport of small hydrophilic molecules across the cytoplasmic membrane. All these proteins share a conserved domain of some two hundred amino acid residues, which includes an ATP-binding site. These proteins are collectively known as ABC transporters. Proteins known to belong to this family are listed below (references are only provided for recently determined sequences). In prokaryotes: -

15

Active transport systems components: alkylphosphonate uptake (phnC/phnK/phnL); arabinose (araG); arginine (artP); dipeptide (dcjAD;dppD/dppF); ferric enterobactin (fepC); ferrichrome (fhuC); galactoside (mglA); glutamine (glnQ); glycerol-3-phosphate (ugpC); glycine betaine/L-proline (proV); glutamate/aspartate (gltL); histidine (hisP); iron(III) (sfuC), iron(III) dicitrate (fecE); lactose (lacK); leucine/isoleucine/valine (braF/braG;livF/livG); maltose (malK); molybdenum (modC); nickel (nikD/nikE); oligopeptide (amiE/amiF;oppD/oppF); peptide (sapD/sapF); phosphate (pstB); putrescine (potG); ribose (rbsA); spermidine/putrescine (potA); sulfate (cysA); vitamin B12 (btuD). -

20

Hemolysin/leukotoxin export proteins hlyB, cyaB and lktB. - Colicin V export protein cvaB. - Lactococcal export protein lcnC [6]. - Lantibiotic transport proteins nisT (nisin) and spaT (subtilin). - Extracellular proteases B and C export protein prtD. - Alkaline protease secretion protein aprD. - Beta-(1,2)-glucan export proteins chvA and ndvA. - Haemophilus influenzae capsule-polysaccharide export protein bexA. - Cytochrome c biogenesis proteins ccmA (also known as cycV and helA). - Polysialic acid transport protein kpsT. - Cell division associated ftsE protein (function unknown). - Copper processing protein nosF from Pseudomonas

25

30

stutzeri. - Nodulation protein nodI from Rhizobium (function unknown). - Escherichia coli proteins cydC and cydD. - Subunit A of the ABC excision nuclease (gene uvrA). - Erythromycin resistance protein from Staphylococcus epidermidis (gene msrA). - Tylosin resistance protein from Streptomyces fradiae (gene tlrC) [7]. - Heterocyst differentiation protein (gene hetA) from Anabaena PCC 7120. - Protein P29 from Mycoplasma hyorhinitis, a probable component of a high affinity transport system. - yhbG, a putative protein whose gene is linked with ntrA in many bacteria such as Escherichia coli, Klebsiella pneumoniae, Pseudomonas putida, Rhizobium meliloti and Thiobacillus ferrooxidans. - Escherichia coli and related bacteria hypothetical proteins yabJ, yadG, yagC, ybbA, ycjW, yddA, yehX, yejF, yheS, yhiG, yhiH, yjcW, yjjK, yojI, yrbF and ytfR. In eukaryotes: - The multidrug transporters (Mdr) (P-glycoprotein), a family of closely related proteins which extrude a wide variety of drugs out of the cell (for a review see [8]). - Cystic fibrosis transmembrane conductance regulator (CFTR), which is most probably involved in the transport of chloride ions. - Antigen peptide transporters 1 (TAP1, PSF1, RING4, HAM-1, mtp1) and 2 (TAP2, PSF2, RING11, HAM-2, mtp2), which are involved in the transport of antigens from the cytoplasm to a membrane-bound compartment for association with MHC class I molecules. - 70 Kd peroxisomal membrane protein (PMP70). - ALDP, a peroxisomal protein involved in X-linked adrenoleukodystrophy [9]. - Sulfonylurea receptor [10], a putative subunit of the B-cell ATP-sensitive potassium channel. - Drosophila proteins white (w) and brown (bw), which are involved in the import of ommatidium screening pigments. - Fungal elongation factor 3 (EF-3). - Yeast STE6 which is responsible for the export of the a-factor pheromone. - Yeast mitochondrial transporter ATM1. - Yeast MDL1 and MDL2. - Yeast SNQ2. - Yeast sporidesmin resistance protein (gene PDR5 or STS1 or YDR1). - Fission yeast heavy metal tolerance protein hmt1. This protein is probably involved in the transport of metal-bound phytochelators. - Fission yeast brefeldin A resistance protein (gene bfr1 or hba2). - Fission yeast leptomycin B resistance protein (gene pmd1). - mbpX, a hypothetical chloroplast protein from Liverwort. - Prestalk-specific protein tagB from slime mold. This protein consists of two domains: a N-terminal subtilase catalytic domain and a C-terminal ABC transporter domain. As a signature pattern for this class of proteins, a conserved region which is located between the 'A' and the 'B' motifs of the ATP-binding site was used.

Consensus pattern: [LIVMFYC SEQ ID NO:6)]-[SA]-[SAPGLVFYKQH SEQ ID NO:7)]-G-[DENQMW SEQ ID NO:8)]-[KRQASPCLIMFW SEQ ID NO:9)]-[KRNQSTAVM SEQ ID

NO:10)]-[KRACLVMS SEQ ID NO:11)]-[LIVMFYPAN SEQ ID NO:12)]-{PHY}-
 [LIVMFWS SEQ ID NO:13)]-[SAGCLIVP SEQ ID NO:14)]-{FYWHP SEQ ID NO:15)}-
 {KRHP SEQ ID NO:16)}-[LIVMFYWSTA SEQ ID NO:17)] The ATP-binding region is
 duplicated in *araG*, *mdl*, *msrA*, *rbsA*, *tlrC*, *uvrA*, *yejF*, *Mdr*'s, *CFTR*, *pmd1* and in *EF-3*. In
 5 some of those proteins, the above pattern only detect one of the two copies of the domain.
 The proteins belonging to this family also contain one or two copies of the ATP-binding
 motifs 'A' and 'B'.

- [1] Higgins C.F., Hyde S.C., Mimmack M.M., Gileadi U., Gill D.R., Gallagher M.P. J.
 10 Bioenerg. Biomembr. 22:571-592(1990).
 [2] Higgins C.F., Gallagher M.P., Mimmack M.M., Pearce S.R. BioEssays 8:111-116(1988).
 [3] Higgins C.F., Hiles I.D., Salmond G.P.C., Gill D.R., Downie J.A., Evans I.J., Holland
 I.B., Gray L., Buckels S.D., Bell A.W., Hermodson M.A. Nature 323:448-450(1986).
 [4] Doolittle R.F., Johnson M.S., Husain I., van Houten B., Thomas D.C., Sancar A. Nature
 15 323:451-453(1986).
 [5] Blight M.A., Holland I.B. Mol. Microbiol. 4:873-880(1990).
 [6] Stoddard G.W., Petzel J.P., van Belkum M.J., Kok J., McKay L.L. Appl. Environ.
 Microbiol. 58:1952-1961(1992).
 [7] Rosteck P.R. Jr., Reynolds P.A., Hershberger C.L. Gene 102:27-32(1991).
 20 [8] Gottesman M.M., Pastan I. J. Biol. Chem. 263:12163-12166(1988).
 [9] Valle D., Gaertner J. Nature 361:682-683(1993).
 [10] Aguilar-Bryan L., Nichols C.G., Wechsler S.W., Clement J.P. IV, Boyd A.E. III,
 Gonzalez G., Herrera-Sosa H., Nguy K., Bryan J., Nelson D.A. Science 268:423-426(1995).

25 4. (ACBP)

Acyl-CoA-binding protein signature

Acyl-CoA-binding protein (ACBP) is a small (10 Kd) protein that binds medium- and long-
 30 chain acyl-CoA esters with very high affinity and may function as an intracellular carrier of
 acyl-CoA esters [1]. ACBP is also known as diazepam binding inhibitor (DBI) or endozepine
 (EP) because of its ability to displace diazepam from the benzodiazepine (BZD) recognition
 site located on the GABA type A receptor. It is therefore possible that this protein also acts as

a neuropeptide to modulate the action of the GABA receptor [2].ACBP is a highly conserved protein of about 90 residues that has been so far found in vertebrates, insects and yeast.

ACBP is also related to the N-terminal section of a probable transmembrane protein of unknown function which has been found in mammals. As a signature pattern, the region that corresponds to residues 19 to 37 in mammalian ACBP was selected.

Consensus pattern: P-[STA]-x-[DEN]-x-[LIVMF SEQ ID NO:2)]-x(2)-[LIVMFY SEQ ID NO:18)]-Y-[GSTA SEQ ID NO:19)]-x-[FY]-K- Q-[STA](2)-x-G-

[1] Rose T.M., Schultz E.R., Todaro G.J. Proc. Natl. Acad. Sci. U.S.A. 89:11287-11291(1992).

[2] Costa E., Guidotti A. Life Sci. 49:325-344(1991).

5. (AIRS)

AIR synthase related proteins

This family includes Hydrogen expression/formation protein HypE, AIR synthases, FGAM synthase and selenide, water dikinase.

6. (AMP-binding)

Putative AMP-binding domain signature

It has been shown [1 to 5] that a number of prokaryotic and eukaryotic enzymes which all probably act via an ATP-dependent covalent binding of AMP to their substrate, share a region of sequence similarity. These enzymes are: - Insects luciferase (luciferin 4-monooxygenase). Luciferase produces light by catalyzing the oxidation of luciferin in presence of ATP and molecular oxygen. - Alpha-aminoacidate reductase from yeast (gene LYS2). This enzyme catalyzes the activation of alpha-aminoacidate by ATP-dependent adenylation and the reduction of activated alpha-aminoacidate by NADPH. - Acetate--CoA ligase (acetyl-CoA synthetase), an enzyme that catalyzes the formation of acetyl-CoA from acetate and CoA. - Long-chain-fatty-acid--CoA ligase, an enzyme that activates long-chain

fatty acids for both the synthesis of cellular lipids and their degradation via beta-oxidation. - 4-coumarate--CoA ligase (4CL), a plant enzyme that catalyzes the formation of 4-coumarate-CoA from 4-coumarate and coenzyme A; the branchpoint reactions between general phenylpropanoid metabolism and pathways leading to various specific end products. -

5 O-succinylbenzoic acid--CoA ligase (OSB-CoA synthetase) (gene *menE*) [6], a bacterial enzyme involved in the biosynthesis of menaquinone (vitamin K2). - 4-Chlorobenzoate--CoA ligase (EC 6.2.1.-) (4-CBA--CoA ligase) [7], a *Pseudomonas* enzyme involved in the degradation of 4-CBA. - Indoleacetate--lysine ligase (IAA-lysine synthetase) [8], an enzyme from *Pseudomonas syringae* that converts indoleacetate to IAA-lysine. - Bile acid-CoA ligase

10 (gene *baiB*) from *Eubacterium* strain VPI 12708 [4]. This enzyme catalyzes the ATP-dependent formation of a variety of C-24 bile acid-CoA. - Crotonobetaine/carnitine-CoA ligase (EC 6.3.2.-) from *Escherichia coli* (gene *caiC*). - L-(alpha-aminoadipyl)-L-cysteinyl-D-valine synthetase (ACV synthetase) from various fungi (gene *acvA* or *pcbAB*). This enzyme catalyzes the first step in the biosynthesis of penicillin and cephalosporin, the formation of

15 ACV from the constituent amino acids. The amino acids seem to be activated by adenylation. It is a protein of around 3700 amino acids that contains three related domains of about 1000 amino acids. - Gramicidin S synthetase I (gene *grsA*) from *Bacillus brevis*. This enzyme catalyzes the first step in the biosynthesis of the cyclic antibiotic gramicidin S, the ATP-dependent racemization of phenylalanine - Tyrocidine synthetase I (gene *tycA*) from

20 *Bacillus brevis*. The reaction carried out by *tycA* is identical to that catalyzed by *grsA* - Gramicidin S synthetase II (gene *grsB*) from *Bacillus brevis*. This enzyme is a multifunctional protein that activates and polymerizes proline, valine, ornithine and leucine. *GrsB* consists of four related domains. - Enterobactin synthetase components E (gene *entE*) and F (gene *entF*) from *Escherichia coli*. These two enzymes are involved in the ATP-

25 dependent activation of respectively 2,3-dihydroxybenzoate and serine during enterobactin (enterochelin) biosynthesis. - Cyclic peptide antibiotic surfactin synthase subunits 1, 2 and 3 from *Bacillus subtilis*. Subunits 1 and 2 contains three related domains while subunit 3 only contains a single domain. - HC-toxin synthetase (gene *HTS1*) from *Cochliobolus carbonum*. This enzyme activates the four amino acids (Pro, L-Ala, D-Ala and 2-amino-9,10-epoxi-8-

30 oxodecanoic acid) that make up HC-toxin, a cyclic tetrapeptide. *HTS1* consists of four related domains. There are also some proteins, whose exact function is not yet known, but which are, very probably, also AMP-binding enzymes. These proteins are: - ORA (octapeptide-repeat antigen), a *Plasmodium falciparum* protein whose function is not known but which shows a

high degree of similarity with the above proteins. - AngR, a *Vibrio anguillarum* protein.

AngR is thought to be a transcriptional activator which modulates the anguibactin (an iron-binding siderophore) biosynthesis gene cluster operon. But it is believed [9], that angR is not a DNA-binding protein, but rather an enzyme involved in the biosynthesis of anguibactin.

5 This conclusion is based on three facts: the presence of the AMP-binding domain; the size of angR (1048 residues), which is far bigger than any bacterial transcriptional protein; and the presence of a probable S-acyl thioesterase immediately downstream of angR. - A

hypothetical protein in mmsB 3'region in *Pseudomonas aeruginosa*. - *Escherichia coli*

hypothetical protein ydiD. - Yeast hypothetical protein YBR041w. - Yeast hypothetical

10 protein YBR222c. - Yeast hypothetical protein YER147c. All these proteins contain a highly conserved region very rich in glycine, serine, and threonine which is followed by a conserved lysine. A parallel can be drawn between this type of domain and the G-x(4)-G-K-[ST] ATP-/GTP-binding 'P-loop' domain or the protein kinases G-x-G-x(2)-[SG]-x(10,20)-KATP-binding domains.

15 Consensus pattern: [LIVMFY SEQ ID NO:18)]-x(2)-[STG]-[STAG SEQ ID NO:20)]-G-[ST]-[STEI SEQ ID NO:21)]-[SG]-x-[PASLIVM SEQ ID NO:22)]- [KR] In a majority of cases the residue that follows the Lys at the end of the pattern is a Gly.

20 [1] Toh H. Protein Seq. Data Anal. 4:111-117(1991).

[2] Smith D.J., Earl A.J., Turner G. EMBO J. 9:2743-2750(1990).

[3] Schroeder J. Nucleic Acids Res. 17:460-460(1989).

[4] Mallonee D.H., Adams J.L., Hylemon P.B. J. Bacteriol. 174:2065-2071(1992).

[5] Turgay K., Krause M., Marahiel M.A. Mol. Microbiol. 6:529-546(1992).

25 [6] Driscoll J.R., Taber H.W. J. Bacteriol. 174:5063-5071(1992).

[7] Babbitt P.C., Kenyon G.L., Matin B.M., Charest H., Sylvestre M., Scholten J.D., Chang K.-H., Liang P.-H., Dunaway-Mariano D. Biochemistry 31:5594-5604(1992).

[8] Farrell D.H., Mikesell P., Actis L.A., Crosa J.H. Gene 86:45-51(1990).

30 7. AP2 domain

This 60 amino acid residue domain can bind to DNA [1]. This domain is plant specific. Members of this family are suggested to be related to pyridoxal phosphate-binding domains such as found in aminotran_2 [3]. AP2 domains are also described in Jofuku et al., co-pending U.S. Patent applications 08/700,152, 08/879,827, 08/912,272, 09/026,039.

5

[1] Ohme-takagi M, Shinshi H; Plant Cell 1995;7:173-182.

[2] Weigel D; Plant Cell 1995;7:388-389.

[3] Mushegian AR, Koonin EV; Genetics 1996;144:817-828.

10

8. ARID

The ARID domain is an AT-Rich Interaction domain sharing structural homology to DNA replication and repair nucleases and polymerases.

15

[1] Herrscher RF, Kaplan MH, Lelsz DL, Das C, Scheuermann R, Tucker PW; Genes Dev 1995;9:3067-3082.

[2] Yuan YC, Whitson RH, Liu Q, Itakura K, Chen Y; Nat Struct Biol 1998;5:959-964.

20

9. (ATP synt)

ATP synthase gamma subunit signature

25

ATP synthase (proton-translocating ATPase) (EC 3.6.1.34) [1,2] is a component of the cytoplasmic membrane of eubacteria, the inner membrane of mitochondria, and the thylakoid membrane of chloroplasts. The ATPase complex is composed of an oligomeric transmembrane sector, called CF(0), and a catalytic core, called coupling factor CF(1). The former acts as a proton channel; the latter is composed of five subunits, alpha, beta, gamma, delta and epsilon. Subunit gamma is believed to be important in regulating ATPase activity and the flow of protons through the CF(0) complex. The best conserved region of the gamma subunit [3] is its C-terminus which seems to be essential for assembly and catalysis. As a signature pattern to detect ATPase gamma subunits, a 14 residue conserved segment where the last amino acid is found one to three residues from the C-terminal extremity was used.

30

Consensus pattern: [IV]-T-x-E-x(2)-[DE]-x(3)-G-A-x-[SAKR SEQ ID NO:23])- Note: Pea chloroplast gamma and two Bacillus species gamma subunits are not detected by this motif.

[1] Futai M., Noumi T., Maeda M. Annu. Rev. Biochem. 58:111-136(1989).

5 [2] Senior A.E. Physiol. Rev. 68:177-231(1988).

[3] Miki J., Maeda M., Mukohata Y., Futai M. FEBS Lett. 232:221-226(1988).

10. (ATP Synt A)

10 Synthase a subunit signature

ATP synthase (proton-translocating ATPase) (EC 3.6.1.34) [1,2] is a component of the cytoplasmic membrane of eubacteria, the inner membrane of mitochondria, and the thylakoid membrane of chloroplasts. The ATPase complex is composed of an oligomeric transmembrane sector, called CF(0), which acts as a proton channel, and a catalytic core, termed coupling factor CF(1). The CF(0) a subunit, also called protein 6, is a key component of the proton channel; it may play a direct role in translocating protons across the membrane. It is a highly hydrophobic protein that has been predicted to contain 8 transmembrane regions [3]. Sequence comparison of a subunits from all available sources reveals very few conserved regions. The best conserved region is located in what is predicted to be the fifth transmembrane domain. This region contains three perfectly conserved residues: an arginine, a leucine and an asparagine. Mutagenesis experiments of ATPase activity. This region was selected as a signature pattern.

25 Consensus pattern: [STAGN SEQ ID NO:24])-x-[STAG SEQ ID NO:20)]-[LIVMF SEQ ID NO:2)]-R-L-x-[SAGV SEQ ID NO:25)]-N-[LIVMT SEQ ID NO:1)] [R is important for proton translocation]

[1] Futai M., Noumi T., Maeda M. Annu. Rev. Biochem. 58:111-136(1989).

30 [2] Senior A.E. Physiol. Rev. 68:177-231(1988).

[3] Lewis M.L., Chang J.A., Simoni R.D. J. Biol. Chem. 265:10541-10550(1990).

[4] Cain B.D., Simoni R.D. J. Biol. Chem. 264:3292-3300(1989).

11. ATP synthase B

Part of the CF(0) (base unit) of the ATP synthase. The base unit is thought to translocate protons through membrane (inner membrane in mitochondria, thylakoid membrane in plants, cytoplasmic membrane in bacteria). The B subunits are thought to interact with the stalk of the CF(1) subunits.

12. (ATP synt C)

ATP synthase c subunit signature

ATP synthase (proton-translocating ATPase) [1,2] is a component of the cytoplasmic membrane of eubacteria, the inner membrane of mitochondria, and the thylakoid membrane of chloroplasts. The ATPase complex is composed of an oligomeric transmembrane sector, called CF(0), which acts as a proton channel, and a catalytic core, termed coupling factor CF(1). The CF(0) c subunit (also called protein 9, proteolipid, or subunit III) [3,4] is a highly hydrophobic protein of about 8 Kd which has been implicated in the proton-conducting activity of ATPase. Structurally subunit c consists of two long terminal hydrophobic regions, which probably span the membrane, and a central hydrophilic region. N,N'-dicyclohexylcarbodiimide (DCCD) can bind covalently to subunit c and thereby abolish the ATPase activity. DCCD binds to a specific glutamate or aspartate residue which is located in the middle of the second hydrophobic region near the C-terminus of the protein. A signature pattern which includes the DCCD-binding residue was derived.

Consensus pattern: [GSTA SEQ ID NO:19]-R-[NQ]-P-x(10)-[LIVMFYW SEQ ID NO:26]](2)-x(3)-[LIVMFYW SEQ ID NO:26]]-x-[DE] [D or E binds DCCD]

[1] Futai M., Noumi T., Maeda M. Annu. Rev. Biochem. 58:111-136(1989).

[2] Senior A.E. Physiol. Rev. 68:177-231(1988).

[3] Ivaschenko A.T., Karpenyuk T.A., Ponomarenko S.V. Biokhimiia 56:406-419(1991).

[4] Recipon H., Perasso R., Adoutte A., Quetier F. J. Mol. Evol. 34:292-303(1992).

13. (ATP synt DE)

ATP synthase, Delta/Epsilon chain

Part of the ATP synthase CF(1). These subunits are part of the head unit of the ATP synthase.

- 5 The subunits are called delta and epsilon in human and metazoan species but in bacterial species the delta (D) subunit is theequivalent to the Oligomycin sensitive subunit (OSCP) in metozoans.

10 14. (ATP synt ab)

ATP synthase alpha and beta subunits signature

- ATP synthase (proton-translocating ATPase) [1,2] is a component of the cytoplasmic membrane of eubacteria, the inner membrane of mitochondria, and the thylakoid membrane of chloroplasts. The ATPase complex is composed of an oligomeric transmembrane sector, called CF(0), and a catalytic core, called coupling factor CF(1). The former acts as a proton channel; the latter is composed of five subunits, alpha, beta, gamma, delta and epsilon. The sequences of subunits alpha and beta are related and both contain a nucleotide-binding site for ATP and ADP. The beta chain has catalytic activity, while the alpha chain is a regulatory subunit. Vacuolar ATPases [3] (V-ATPases) are responsible for acidifying a variety of intracellular compartments in eukaryotic cells. Like F-ATPases, they are oligomeric complexes of a transmembrane and a catalytic sector. The sequence of the largest subunit of the catalytic sector (70 Kd) is related to that of F-ATPase beta subunit, while a 60 Kd subunit, from the same sector, is related to the F-ATPases alpha subunit [4]. Archaeobacterial membrane-associated ATPases are composed of three subunits. The alpha chain is related to F-ATPases beta chain and the beta chain is related to F-ATPases alpha chain [4]. A protein highly similar to F-ATPase beta subunits is found [5] in some bacterial apparatus involved in a specialized protein export pathway that proceeds without signal peptide cleavage. This protein is known as flil in *Bacillus* and *Salmonella*, Spa47 (mxlB) in *Shigella flexneri*, HrpB6 in *Xanthomonas campestris* and yscN in *Yersinia* virulence plasmids. To detect these ATPase subunits, a segment of ten amino-acid residues, containing two conserved serines, as a signature pattern was selected. The first serine seems to be important for catalysis - in the ATPase alpha chain at least - as its mutagenesis causes catalytic impairment.
- 15
- 20
- 25
- 30

Consensus pattern: P-[SAP]-[LIV]-[DNH]-x(3)-S-x-S [The first S is a putative active site residue]

- 5 [1] Futai M., Noumi T., Maeda M. *Annu. Rev. Biochem.* 58:111-136(1989).
[2] Senior A.E. *Physiol. Rev.* 68:177-231(1988).
[3] Nelson N. J. *Bioenerg. Biomembr.* 21:553-571(1989).
[4] Gogarten J.P., Kibak H., Dittrich P., Taiz L., Bowman E.J., Bowman B.J., Manolson
M.F., Poole R.J., Date T., Oshima T., Konishi J., Denda K., Yoshida M. *Proc. Natl. Acad.*
10 *Sci. U.S.A.* 86:6661-6665(1989).
[5] Dreyfus G., Williams A.W., Kawagishi I., MacNab R.M. *J. Bacteriol.* 175:3131-
3138(1993).

- 15 15. (ATP synt ab C)
ATP synthase ab C terminal.

Number of members: 190

- 20 [1] Abrahams JP, Leslie AG, Lutter R, Walker JE; "Structure at 2.8 Å resolution of F1-
ATPase from bovine heart mitochondria." *Nature* 1994;370:621-628.

16. (A deaminase)
25 Adenosine and AMP deaminase signature

- Adenosine deaminase catalyzes the hydrolytic deamination of adenosine into inosine. AMP
deaminase catalyzes the hydrolytic deamination of AMP into IMP. It has been shown [1] that
these two types of enzymes share three regions of sequence similarities; these regions are
30 centered on residues which are proposed to play an important role in the catalytic mechanism
of these two enzymes. One of these regions, containing two conserved aspartic acid residues
that are potential active site residues was selected.

Consensus pattern: [SA]-[LIVM SEQ ID NO:4)]-[NGS]-[STA]-D-D-P [The two D's are putative active site residues]

[1] Chang Z., Nygaard P., Chinault A.C., Kellems R.E. Biochemistry 30:2273-2280(1991).

5

17. (Acetyltransf)

Acetyltransferase (GNAT) family.

10 This family contains proteins with N-acetyltransferase functions.

[1] Neuwald AF, Landsman D; Trends Biochem Sci 1997;22:154-155.

15 18. (Aconitase C)

Aconitase family signature

Aconitase (aconitate hydratase) (EC 4.2.1.3) [1] is the enzyme from the tricarboxylic acid cycle that catalyzes the reversible isomerization of citrate and isocitrate. Cis-aconitate is
20 formed as an intermediary product during the course of the reaction. In eukaryotes two isozymes of aconitase are known to exist: one found in the mitochondrial matrix and the other found in the cytoplasm. Aconitase, in its active form, contains a 4Fe-4S iron-sulfur cluster; three cysteine residues have been shown to be ligands of the 4Fe-4S cluster. It has been shown that the aconitase family also contains the following proteins: - Iron-responsive
25 element binding protein (IRE-BP). IRE-BP is a cytosolic protein that binds to iron-responsive elements (IREs). IREs are stem-loop structures found in the 5'UTR of ferritin, and delta aminolevulinic acid synthase mRNAs, and in the 3'UTR of transferrin receptor mRNA. IRE-BP also express aconitase activity. - 3-isopropylmalate dehydratase (EC 4.2.1.33) (isopropylmalate isomerase), the enzyme that catalyzes the second step in the biosynthesis of
30 leucine. - Homoaconitase (EC 4.2.1.36) (homoaconitate hydratase), an enzyme that participates in the alpha-amino adipate pathway of lysine biosynthesis and that converts cis-homoaconitate into homoisocitric acid. - Escherichia coli protein ybhJ. As a signature for

proteins from the aconitase family, two conserved regions that contain the three cysteine ligands of the 4Fe-4S cluster were selected.

Consensus pattern: [LIVM SEQ ID NO:4)]-x(2)-[GSACIVM SEQ ID NO:27)]-x-[LIV]-
 5 [GTIV SEQ ID NO:28)]-[STP]-C-x(0,1)-T-N- [GSTANI SEQ ID NO:29)]-x(4)-[LIVMA
 SEQ ID NO:30)] [C binds the iron-sulfur center]

Consensus pattern: G-x(2)-[LIVWPQ SEQ ID NO:31)]-x(3)-[GAC]-C-[GSTAM SEQ ID
 NO:32)]-[LIMPTA SEQ ID NO:33)]-C-[LIMV SEQ ID NO:34)]- [GA] [The two C's bind
 10 the iron-sulfur center]

[1] Gruer M.J., Artymiuk P.J., Guest J.R. Trends Biochem. Sci. 22:3-6(1997).

19. (Acyl-CoA dh)

Acyl-CoA dehydrogenases signatures

Acyl-CoA dehydrogenases [1,2,3] are enzymes that catalyze the alpha, beta-dehydrogenation of acyl-CoA esters and transfer electrons to ETF, the electron transfer protein. Acyl-CoA
 20 dehydrogenases are FAD flavoproteins. This family currently includes: - Five eukaryotic isozymes that catalyze the first step of the beta-oxidation cycles for fatty acids with various chain lengths. These are short (SCAD) (EC 1.3.99.2), medium (MCAD) (EC 1.3.99.3), long (LCAD) (EC 1.3.99.13), very-long (VLCAD) and short/branched (SBCAD) chain acyl-CoA dehydrogenases. These enzymes are located in the mitochondrion. They are all
 25 homotetrameric proteins of about 400 amino acid residues except VLCAD which is a dimer and which contains, in its mature form, about 600 residues. - Glutaryl-CoA dehydrogenase (EC 1.3.99.7) (GCDH), which is involved in the catabolism of lysine, hydroxylysine and tryptophan. - Isovaleryl-CoA dehydrogenase (EC 1.3.99.10) (IVD), involved in the catabolism of leucine. - Acyl-coA dehydrogenases acsA and mmgC from *Bacillus subtilis*. -
 30 Butyryl-CoA dehydrogenase (EC 1.3.99.2) from *Clostridium acetobutylicum*. - *Escherichia coli* protein caiA [4]. - *Escherichia coli* protein aidB. Two conserved regions were selected as signature patterns. The first is located in the center of these enzymes, the second in the C-terminal section.

Consensus pattern: [GAC]-[LIVM SEQ ID NO:4)]-[ST]-E-x(2)-[GSAN SEQ ID NO:35)]-G-[ST]-D-x(2)-[GSA]

5 Consensus pattern: [QDE]-x(2)-G-[GS]-x-G-[LIVMFY SEQ ID NO:18)]-x(2)-[DEN]-x(4)-[KR]-x(3)- [DEN]

[1] Tanaka K., Ikeda, Matsubara Y., Hyman D.B. Enzyme 38:91-107(1987).

10 [2] Matsubara Y., Indo Y., Naito E., Ozasa H., Glassberg R., Vockley J., Ikeda Y., Kraus J., Tanaka K. J. Biol. Chem. 264:16321-16331(1989).

[3] Aoyama T., Ueno I., Kamijo T., Hashimoto T. J. Biol. Chem. 269:19088-19094(1994).

[4] Eichler K., Bourgis F., Buchet A., Kleber H.-P., Mandrand-Berthelot M.-A. Mol. Microbiol. 13:775-786(1994).

15

20. (Acyl transf)

Acyl transferase domain

Number of members: 161

20

[1] Serre L, Verbree EC, Dauter Z, Stuitje AR, Derewenda ZS; Medline: [95286570](#) "The Escherichia coli malonyl-CoA:acyl carrier protein transacylase at 1.5-A resolution. Crystal structure of a fatty acid synthase component." J Biol Chem 1995;270:12961-12964.

25

21. Acylphosphatase signatures

30 Acylphosphatase (EC [3.6.1.7](#)) [1,2] catalyzes the hydrolysis of various acylphosphate carboxyl-phosphate bonds such as carbamyl phosphate, succinylphosphate, 1,3-diphosphoglycerate, etc. The physiological role of this enzyme is not yet clear.

Acylphosphatase is a small protein of around 100 amino-acid residues. There are two known isozymes. One seems to be specific to muscular tissues, the other, called 'organ-common type', is found in many different tissues. While acylphosphatase have been so far only

characterized in vertebrates, there are a number of bacterial and archeobacterial hypothetical proteins that are highly similar to that enzyme and that probably possess the same activity. These proteins are: - *Escherichia coli* hypothetical protein yccX. - *Bacillus subtilis* hypothetical protein yfIL. - *Archaeoglobus fulgidus* hypothetical protein AF0818. Two conserved regions were selected as signature patterns. The first is located in the N-terminal section, while the second is found in the central part of the protein sequence.

Consensus pattern: [LIV]-x-G-x-V-Q-G-V-x-[FM]-R

Consensus pattern: G-[FYW]-[AVC]-[KRQAM SEQ ID NO:36)]-N-x(3)-G-x-V-x(5)-G

[1] Stefani M., Ramponi G. Life Chem. Rep. 12:271-301(1995).

[2] Stefani M., Taddei N., Ramponi G. Cell. Mol. Life Sci. 53:141-151(1997).

22. (Adap comp sub)

Clathrin adaptor complexes medium chain signatures.

Clathrin coated vesicles (CCV) mediate intracellular membrane traffic such as receptor mediated endocytosis. In addition to clathrin, the CCV are composed of a number of other components including oligomeric complexes which are known as adaptor or clathrin assembly proteins (AP) complexes [1]. The adaptor complexes are believed to interact with the cytoplasmic tails of membrane proteins, leading to their selection and concentration. In mammals two type of adaptor complexes are known: AP-1 which is associated with the Golgi complex and AP-2 which is associated with the plasma membrane. Both AP-1 and AP-2 are heterotetramers that consist of two large chains - the adaptins - (gamma and beta' in AP-1; alpha and beta in AP-2); a medium chain (AP47 in AP-1; AP50 in AP-2) and a small chain (AP19 in AP-1; AP17 in AP-2). The medium chains of AP-1 and AP-2 are evolutionary related proteins of about 50 Kd. Homologs of AP47 and AP50 have also been found in *Caenorhabditis elegans* (genes unc-101 and ap50) [2] and yeast (gene APM1 or YAP54) [3]. Some more divergent, but clearly evolutionary related proteins have also been found in yeast: APM2 and YBR288c. Two conserved regions were selected as signature patterns, one located in the N-terminal region, the other from the central section of these proteins.

Consensus pattern: [IVT]-[GSP]-W-R-x(2,3)-[GAD]-x(2)-[HY]-x(2)-N-x- [LIVMAFY SEQ ID NO:37)](3)-D-[LIVM SEQ ID NO:4)]-[LIVMT SEQ ID NO:1)]-E

5 Consensus pattern: [LIV]-x-F-I-P-P-x-G-x-[LIVMFY SEQ ID NO:18)]-x-L-x(2)-Y

[1] Pearse B.M., Robinson M.S. Annu. Rev. Cell Biol. 6:151-171(1990).

[2] Lee J., Jongeward G.D., Sternberg P.W. Genes Dev. 8:60-73(1994).

[3] Nakayama Y., Goebel M., O'Brine G.B., Lemmon S., Pingchang C.E., Kirchhausen T.
10 Eur. J. Biochem. 202:569-574(1991).

23. (Adenylosucc synt)

Adenylosuccinate synthetase signatures

15

Adenylosuccinate synthetase (EC 6.3.4.4) [1] plays an important role in purinebiosynthesis, by catalyzing the GTP-dependent conversion of IMP and aspartic acid to AMP.

Adenylosuccinate synthetase has been characterized from various sources ranging from Escherichia coli (gene purA) to vertebrate tissues. Invertebrates, two isozymes are present -
20 one involved in purine biosynthesis and the other in the purine nucleotide cycle. Two conserved regions were selected as signature patterns. The first one is a perfectly conserved octapeptide located in the N-terminal section and which is involved in GTP-binding [2]. The second one includes a lysine residue known [2] to be essential for the enzyme's activity.

25 Consensus pattern: Q-W-G-D-E-G-K-G

Consensus pattern: G-I-[GR]-P-x-Y-x(2)-K-x(2)-R [K is the active site residue]

[1] Wiesmueller L., Wittbrodt J., Noegel A.A., Schleicher M. J. Biol. Chem. 266:2480-
30 2485(1991).

[2] Silva M.M., Poland B.W., Hoffman C.R., Fromm H.J., Honzatko R.B. J. Mol. Biol. 254:431-446(1995).

[3] Bouyoub A., Barbier G., Forterre P., Labedan B. 2.3.CO;2-"J. Mol. Biol. 261:144-154(1996).

5 24. (AdoHcyase)

S-adenosyl-L-homocysteine hydrolase signatures

S-adenosyl-L-homocysteine hydrolase (EC 3.3.1.1) (AdoHcyase) is an enzyme of the activated methyl cycle, responsible for the reversible hydration of S-adenosyl-L-homocysteine into adenosine and homocysteine. AdoHcyase is an ubiquitous enzyme which binds and requires NAD⁺ as a cofactor. AdoHcyase is a highly conserved protein [1] of about 430 to 470 amino acids. Two highly conserved regions were selected as signature patterns. The first pattern is located in the N-terminal section; the second is derived from a glycine-rich region in the central part of AdoHcyase; a region thought to be involved in NAD-binding.

15 Consensus pattern: [GSA]-[CS]-N-x-[FYLM SEQ ID NO:38)]-S-[ST]-[QA]-[DEN]-x-[AV]-[AT]-[AD]-[AC]-[LIVMCG SEQ ID NO:39)]

20 Consensus pattern: [GA]-[KS]-x(3)-[LIV]-x-G-[FY]-G-x-[VC]-G-[KRL]-G-x-[ASC]

[1] Sganga M.W., Aksamit R.R., Cantoni G.L., Bauer C.E. Proc. Natl. Acad. Sci. U.S.A. 89:6328-6332(1992).

25 25. AhpC/TSA family

This family contains proteins related to alkyl hydroperoxide reductase. Comment: (AhpC) and thiol specific antioxidant (TSA).

30 [1] Chae HZ, Robison K, Poole LB, Church G, Storz G, Rhee SG, Proc Natl Acad Sci U S A 1994;91:7017-7021

26. (Aldose epim)

Aldose 1-epimerase putative active site Aldose 1-epimerase (EC 5.1.3.3) (mutarotase) is the enzyme responsible for the anomeric interconversion of D-glucose and other aldoses between their alpha- and beta-forms. The sequence of mutarotase from two bacteria, *Acinetobacter calcoaceticus* and *Streptococcus thermophilus* is available [1]. It has also been shown that, on the basis of extensive sequence similarities, a mutarotase domain seem to be present in the C-terminal half of the fungal GAL10 protein which encodes, in the N-terminal part, for UDP-glucose 4-epimerase. The best conserved region in the sequence of mutarotase is centered around a conserved histidine residue which may be involved in the catalytic mechanism.

Consensus pattern: [NS]-x-T-N-H-x-Y-[FW]-N-[LI]

[1] Poolman B., Royer T.J., Mainzer S.E., Schmidt B.F. J. Bacteriol. 172:4037-4047(1990).

27. (AlkA DNA repair)

Alkylbase DNA glycosidases alkA family signature

Alkylbase DNA glycosidases [1] are DNA repair enzymes that hydrolyzes the deoxyribose N-glycosidic bond to excise various alkylated bases from a damaged DNA polymer. In *Escherichia coli* there are two alkylbase DNA glycosidases: one (gene tag) which is constitutively expressed and which is specific for the removal of 3-methyladenine (EC 3.2.2.20), and one (gene alkA) which is induced during adaptation to alkylation and which can remove a variety of alkylation products (EC 3.2.2.21). Tag and alkA do not share any region of sequence similarity. In yeast there is an alkylbase DNA glycosidase (gene MAG1) [2,3], which can remove 3-methyladenine or 7-methyladenine and which is structurally related to alkA. MAG and alkA are both proteins of about 300 amino acid residues. While the C- and N-terminal ends appear to be unrelated, there is a central region of about 130 residues which is well conserved. A portion of this region has been selected as a signature pattern .

Consensus pattern: G-I-G-x-W-[ST]-[AV]-x-[LIVMFY SEQ ID NO:18]](2)-x-[LIVM SEQ ID NO:4)]-x(8)-[MF]-x(2)- [ED]-D

[1] Lindahl T., Sedgwick B. Annu. Rev. Biochem. 57:133-157(1988).

5 [2] Berdal K.G., Bjoras M., Bjelland S., Seeberg E.C. EMBO J. 9:4563-4568(1990).

[3] Chen J., Derfler B., Samson L. EMBO J. 9:4569-4575(1990).

28. Ammonium transporters signature

10

A number of proteins involved in the transport of ammonium ions across a membrane as well as some yet uncharacterized proteins have been shown [1,2] to be evolutionary related. These proteins are: - Yeast ammonium transporters MEP1, MEP2 and MEP3. - Arabidopsis thaliana high affinity ammonium transporter (gene AMT1). - Corynebacterium glutamicum ammonium and methylammonium transport system. - Escherichia coli putative ammonium transporter amtB. - Bacillus subtilis nrgA. - Mycobacterium tuberculosis hypothetical protein MtCY338.09c. - Synechocystis strain PCC 6803 hypothetical proteins sl10108, sl10537 and sl10117. - Methanococcus jannaschii hypothetical proteins MJ0058 and MJ1343. - Caenorhabditis elegans hypothetical proteins C05E11.4, F49E11.3 and M195.3. As
 15
 20 expected by their transport function, these proteins are highly hydrophobic and seem to contain from 10 to 12 transmembrane domains. The best conserved region seems to be located in the fifth (or sixth) transmembrane region and is used as a signature pattern.

Consensus pattern: D-[FYWS SEQ ID NO:40)]-A-G-[GSC]-x(2)-[IV]-x(3)-[SAG](2)-x(2)-
 25 [SAG]- [LIVMF SEQ ID NO:2)]-x(3)-[LIVMFYWA SEQ ID NO:41]](2)-x-[GK]-x-R

[1] Ninnemann O., Janniaux J.-C., Frommer W.B. EMBO J. 13:3464-3471(1994).

[2] Siewe R.M., Weil B., Burkovski A., Eikmanns B.J., Eikmanns M., Kraemer R. J. Biol. Chem. 271:5398-5403(1996).

30 [3] Saier M.H. Jr. Adv. Microbiol. Physiol. 40:81-136(1998).

29. (Arch_histone)

CBF/NF-Y subunits signatures

Diverse DNA binding proteins are known to bind the CCAAT box, a common cis-acting element found in the promoter and enhancer regions of a large number of genes in eukaryotes. Amongst these proteins is one known as the CCAAT-binding factor (CBF) or NF-Y [1]. CBF is a heteromeric transcription factor that consists of two different components both needed for DNA-binding. The HAP protein complex of yeast binds to the upstream activation site of cytochrome C iso-1 gene (CYC1) as well as other genes involved in mitochondrial electron transport and activates their expression. It also recognizes the sequence CCAAT and is structurally and evolutionary related to CBF. The first subunit of CBF, known as CBF-A or NF-YB in vertebrates, HAP3 in budding yeast and as php3 in fission yeast, is a protein of 116 to 210 amino-acid residues which contains a highly conserved central domain of about 90 residues. This domain seems to be involved in DNA-binding; a signature pattern had been developed from its central part. The second subunit of CBF, known as CBF-B or NF-YA in vertebrates, HAP2 in budding yeast and php2 in fission yeast, is a protein of 265 to 350 amino-acid residues which contains a highly conserved region of about 60 residues. This region, called the 'essential core' [2], seems to consist of two subdomains: an N-terminal subunit-association domain and a C-terminal DNA recognition domain. A signature pattern has been developed from a section of the subunit-association domain.

Consensus pattern: C-V-S-E-x-I-S-F-[LIVM SEQ ID NO:4]-T-[SG]-E-A-[SC]-[DE]-[KRQ]-C-

Consensus pattern: Y-V-N-A-K-Q-Y-x-R-I-L-K-R-R-x-A-R-A-K-L-E-

[1] Li X.-Y., Mantovani R., Hooft van Huijsduijnen R., Andre I., Benoist C., Mathis D. Nucleic Acids Res. 20:1087-1091(1992).

[2] Olesen J.T., Fikes J.D., Guarente L. Mol. Cell. Biol. 11:611-619(1991).

30. Argininosuccinate synthase signatures

Argininosuccinate synthase (EC 6.3.4.5) (AS) is a urea cycle enzyme that catalyzes the penultimate step in arginine biosynthesis: the ATP-dependent ligation of citrulline to aspartate to form argininosuccinate, AMP and pyrophosphate [1,2]. In humans, a defect in the AS gene causes citrullinemia, a genetic disease characterized by severe vomiting spells and mental retardation. AS is a homotetrameric enzyme of chains of about 400 amino-acid residues. An arginine seems to be important for the enzyme's catalytic mechanism. The sequences of AS from various prokaryotes, archaeobacteria and eukaryotes show significant similarity. Two signature patterns have been selected for AS. The first is a highly conserved stretch of nine residues located in the N-terminal extremity of these enzymes, the second is derived from a conserved region which contains one of the conserved arginine residues.

Consensus pattern: [AS]-[FY]-S-G-G-[LV]-D-T-[ST]-

Consensus pattern: G-x-T-x-K-G-N-D-x(2)-R-F-

[1] van Vliet F., Crabeel M., Boyen A., Tricot C., Stalon V., Falmagne P., Nakamura Y., Baumberg S., Glansdorff N. Gene 95:99-104(1990).

[2] Morris C.J., Reeve J.N. J. Bacteriol. 170:3125-3130(1988).

31. Armadillo/beta-catenin-like repeats

Approx. 40 amino acid repeat. Tandem repeats form super-helix of helices that is proposed to mediate interaction of beta-catenin with its ligands. CAUTION: This family does not contain all known armadillo repeats.

[1] Huber AH, Nelson WJ, Weis WI, Cell 1997;90:871-882.

[2] Gumbiner BM, Curr Opin Cell Biol 1995;7:634-640.

[3] Cavallo R, Rubenstein D, Peifer M, Curr Opin Genet Dev 1997;7:459-466.

[4] Su LK, Vogelstein B, Kinzler KW, Science 1993;262:1734-1737.

[5] Masiarz FR, Munemitsu S, Polakis P Science 1993;262:1731-1734

[6] Peifer M, Wieschaus E, Cell 1990;63:1167-1176.

32. (Asn Synthase)

Asparagine synthase

This family is always found associated with GATase_2. Members of this family catalyse the conversion of aspartate to asparagine.

33. Asparaginase_2

Asparaginase 12 members

34. (Aspartyl tRNA N)

Aminoacyl-transfer RNA synthetases class-II signatures

Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure. The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6] and probably have a common folding pattern in their catalytic domain for the binding of ATP and amino acid which is different to the Rossmann fold observed for the class I synthetases [7]. Class-II tRNA synthetases do not share a high degree of similarity, however at least three conserved regions are present [2,5,8]. Signature patterns have been derived from two of these regions.

Consensus pattern: [FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE]

Consensus pattern: [GSTALVF SEQ ID NO:42)]-{DENQHRKP SEQ ID NO:43)}-[GSTA SEQ ID NO:19)]-[LIVMF SEQ ID NO:2)]-[DE]-R-[LIVMF SEQ ID NO:2)]-x-[LIVMSTAG SEQ ID NO:44)]-[LIVMFY SEQ ID NO:18)]

- [1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).
[2] Delarue M., Moras D. BioEssays 15:675-687(1993).
[3] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).
[4] Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).
5 [5] Cusack S., Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).
[6] Cusack S. Biochimie 75:1077-1081(1993).
[7] Cusack S., Berthet-Colominas C., Haertlein M., Nassar N., Leberman R. Nature 347:249-255(1990).
[8] Leveque F., Plateau P., Dessen P., Blanquet S. Nucleic Acids Res. 18:305-312(1990).

10

35. (ArfGap) Putative GTP-ase activating protein for Arf. Putative zinc fingers with GTPase activating proteins (GAPs) towards the small GTPase, Arf. The GAP of ARD1 stimulates GTPase hydrolysis for ARD1 but not ARFs. Number of members: 34

15

[1]Medline: 96324970. Identification and cloning of centaurin-alpha. A novel phosphatidylinositol 3,4,5-trisphosphate-binding protein from rat brain. Hammonds-Odie LP, Jackson TR, Profit AA, Blader IJ, Turck CW, Prestwich GD, Theibert AB; J Biol Chem 1996;271:18859-18868.

20

[2]Medline: 97296423. A target of phosphatidylinositol 3,4,5-trisphosphate with a zinc finger motif similar to that of the ADP-ribosylation -factor GTPase-activating protein and two pleckstrin homology domains. Tanaka K, Imajoh-Ohmi S, Sawada T, Shirai R, Hashimoto Y, Iwasaki S, Kaibuchi K, Kanaho Y, Shirai T, Terada Y, Kimura K, Nagata S, Fukui Y; Eur J Biochem 1997;245:512-519.

25

[3] 98112795. Molecular characterization of the GTPase-activating domain of ADP-ribosylation factor domain protein 1 (ARD1). Vitale N, Moss J, Vaughan M; J Biol Chem 1998;273:2553-2560.

30

36. Apolipoprotein. Apolipoprotein A1/A4/E family. This family includes: Swiss:P02647 Apolipoprotein A-I. Swiss:P06727 Apolipoprotein A-IV. Swiss:P02649 Apolipoprotein E. These proteins contain several 22 residue repeats which form a pair of alpha helices. Number of members: 42

[1]Medline: 91289138. Three-dimensional structure of the LDL receptor-binding domain of human apolipoprotein E. Wilson C, Wardell MR, Weisgraber KH, Mahley RW, Agard DA; Science 1991;252:1817-1822.

5

37. Amino acid permeases signature

Amino acid permeases are integral membrane proteins involved in the transport of amino acids into the cell. A number of such proteins have been found to be evolutionary related [1,2,3]. These proteins are: - Yeast general amino acid permeases (genes GAP1, AGP2 and AGP3). - Yeast basic amino acid permease (gene ALP1). - Yeast Leu/Val/Ile permease (gene BAP2). - Yeast arginine permease (gene CAN1). - Yeast dicarboxylic amino acid permease (gene DIP5). - Yeast asparagine/glutamine permease (gene AGP1). - Yeast glutamine permease (gene GNP1). - Yeast histidine permease (gene HIP1). - Yeast lysine permease (gene LYP1). - Yeast proline permease (gene PUT4). - Yeast valine and tyrosine permease (gene VAL1/TAT1). - Yeast tryptophan permease (gene TAT2/SCM2). - Yeast choline transport protein (gene HNM1/CTR1). - Yeast GABA permease (gene UGA4). - Yeast hypothetical protein YKL174c. - Fission yeast protein isp5. - Fission yeast hypothetical protein SpAC8A4.11 - Fission yeast hypothetical protein SpAC11D3.08c. - *Emmericella nidulans* proline transport protein (gene prnB). - *Trichoderma harzianum* amino acid permease INDA1. - *Salmonella typhimurium* L-asparagine permease (gene ansP). - *Escherichia coli* aromatic amino acid transport protein (gene aroP). - *Escherichia coli* D-serine/D-alanine/glycine transporter (gene cycA). - *Escherichia coli* GABA permease (gene gabP). - *Escherichia coli* lysine-specific permease (gene lysP). - *Escherichia coli* phenylalanine-specific permease (gene pheP). - *Salmonella typhimurium* proline-specific permease (gene proY). - *Escherichia coli* and *Klebsiella pneumoniae* hypothetical protein yeeF. - *Escherichia coli* and *Salmonella typhimurium* hypothetical protein yifK. - *Bacillus subtilis* permeases rocC and rocE which probably transports arginine or ornithine. These proteins seem to contain up to 12 transmembrane segments. As a signature for this family of proteins, the best conserved region which is located in the second transmembrane segment has been selected.

Consensus pattern: [STAGC SEQ ID NO:45)]-G-[PAG]-x(2,3)-[LIVMFYWA SEQ ID NO:41)](2)-x-[LIVMFYW SEQ ID NO:26)]-x- [LIVMFWSTAGC SEQ ID NO:46)](2)-[STAGC SEQ ID NO:45)]-x(3)-[LIVMFYWT SEQ ID NO:47)]-x-[LIVMST SEQ ID NO:48)]-x(3)- [LIVMCTA SEQ ID NO:49)]-[GA]-E-x(5)-[PSAL SEQ ID NO:50)]-

- [1] Weber E., Chevalier M.R., Jund R. J. Mol. Evol. 27:341-350(1988).
 [2] Vandenbol M., Jauniaux J.-C., Grenson M. Gene 83:153-159(1989).
 [3] Reizer J., Finley K., Kakuda D., McLeod C.L., Reizer A., Saier M.H. Jr. Protein Sci. 2:20-30(1993).

38. aakinase (1) Glutamate 5-kinase signature

Glutamate 5-kinase (EC 2.7.2.11) (gamma-glutamyl kinase) (GK) is the enzyme that catalyzes the first step in the biosynthesis of proline from glutamate, the ATP-dependent phosphorylation of L-glutamate into L-glutamate 5-phosphate. In eubacteria (gene proB) and yeast [1] (gene PRO1), GK is a monofunctional protein, while in plants and mammals, it is a bifunctional enzyme (P5CS) [2] that consists of two domains: a N-terminal GK domain and a C-terminal gamma-glutamyl phosphate reductase domain (EC 1.2.1.41) (see <PDOC00940>). As a signature pattern, a highly conserved glycine-and alanine-rich region located in the central section of these enzymes has been selected. Yeast hypothetical protein YHR033w is highly similar to GK.

Consensus pattern: [GSTN SEQ ID NO:51)]-x(2)-G-x-G-[GC]-[IM]-x-[STA]-K-[LIVM SEQ ID NO:4)]-x-[SA]-[TCA]- x(2)-[GALV SEQ ID NO:52)]-x(3)-G-

- [1] Li W., Brandriss M.C. J. Bacteriol. 174:4148-4156(1992).
 [2] Hu C.-A.A., Delauney A.J., Verma D.P.S. Proc. Natl. Acad. Sci. U.S.A. 89:9354-9358(1992).

aakinase (2) Aspartokinase signature

Aspartokinase (EC 2.7.2.4) (AK) [1] catalyzes the phosphorylation of aspartate. The product of this reaction can then be used in the biosynthesis of lysine or in the pathway leading to homoserine, which participates in the biosynthesis of threonine, isoleucine and methionine. In

Escherichia coli, there are three different isozymes which differ in their sensitivity to repression and inhibition by Lys, Met and Thr. AK1 (gene thrA) and AK2 (gene metL) are bifunctional enzymes which both consist of an N- terminal AK domain and a C-terminal homoserine dehydrogenase domain. AK1 is involved in threonine biosynthesis and AK2, in that of methionine. The third isozyme, AK3 (gene lysC), is monofunctional and involved in lysine synthesis. In yeast, there is a single isozyme of AK (gene HOM3). As a signature pattern for AK, a conserved region located in the N-terminal extremity has been selected.

Consensus pattern: [LIVM SEQ ID NO:4)]-x-K-[FY]-G-G-[ST]-[SC]-[LIVM SEQ ID NO:4)]-

[1] Rafalski J.A., Falco S.C. J. Biol. Chem. 263:2146-2151(1988).

aakinase (3) Gamma-glutamyl phosphate reductase signature

Gamma-glutamyl phosphate reductase (EC 1.2.1.41) (GPR) is the enzyme that catalyzes the second step in the biosynthesis of proline from glutamate, the NADP-dependent reduction of L-glutamate 5-phosphate into L-glutamate 5-semialdehyde and phosphate. In eubacteria (gene proA) and yeast [1] (gene PRO2), GPR is a monofunctional protein, while in plants and mammals, it is a bifunctional enzyme (P5CS) [2] that consists of two domains: a N-terminal glutamate 5-kinase domain (EC 2.7.2.11) (see <PDOC00701>) and a C-terminal GPR domain. As a signature pattern, a conserved region that contains two histidine residues has been selected. This region is located in the last third of GPR.

Consensus pattern: V-x(5)-A-[LIV]-x-H-I-x(2)-[HY]-[GS]-[ST]-x-H-[ST]-[DE]-x- I-

[1] Pearson B.M., Hernando Y., Payne J., Wolf S.S., Kalogeropoulos A., Schweizer M. Yeast 12:1021-1031(1996).

[2] Hu C.-A.A., Delauney A.J., Verma D.P.S. Proc. Natl. Acad. Sci. U.S.A. 89:9354-9358(1992).

39. (abhydrolase) alpha/beta hydrolase fold. This catalytic domain is found in a very wide range of enzymes.

[1] Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, Harel M, Remington SJ, Silman I, Schrag J, Sussman JL, Verschueren KHG, Goldman A, Protein Eng 1992;5:197-211.

5

40. (Acid phosphatase) Histidine acid phosphatases signatures

Acid phosphatases (EC 3.1.3.2) are a heterogeneous group of proteins that hydrolyze phosphate esters, optimally at low pH. It has been shown [1] that a number of acid phosphatases, from both prokaryotes and eukaryotes, share two regions of sequence similarity, each centered around a conserved histidine residue. These two histidines seem to be involved in the enzymes' catalytic mechanism [2,3]. The first histidine is located in the N-terminal section and forms a phosphohistidine intermediate while the second is located in the C-terminal section and possibly acts as proton donor. Enzymes belonging to this family are called 'histidine acid phosphatases' and are listed below:

10

15

- Escherichia coli pH 2.5 acid phosphatase (gene appA).

- Escherichia coli glucose-1-phosphatase (EC 3.1.3.10) (gene agp).

- Yeast constitutive and repressible acid phosphatases (genes PHO3 and PHO5).

20

- Fission yeast acid phosphatase (gene pho1).

- Aspergillus phytases A and B (EC 3.1.3.8) (gene phyA and phyB).

- Mammalian lysosomal acid phosphatase.

- Mammalian prostatic acid phosphatase.

- Caenorhabditis elegans hypothetical proteins B0361.7, C05C10.1, C05C10.4

25

- and F26C11.1.

Consensus pattern[LIVM SEQ ID NO:4)]-x(2)-[LIVMA SEQ ID NO:30)]-x(2)-[LIVM SEQ ID NO:4)]-x-R-H-[GN]-x-R-x-[PAS] [H is the phosphohistidine residue]

30

Consensus pattern[LIVMF SEQ ID NO:2)]-x-[LIVMFAG SEQ ID NO:53)]-x(2)-[STAGI SEQ ID NO:54)]-H-D-[STANQ SEQ ID NO:55)]-x-[LIVM SEQ ID NO:4)]-x(2)-[LIVMFY SEQ ID NO:18)]-x(2)-[STA] [H is an active site residue] Sequences known to belong to this

class detected by the patternALL, except for rat prostatic acid phosphatase which seems to have Tyr instead of the active site His

[1] van Etten R.L., Davidson R., Stevis P.E., MacArthur H., Moore D.L. J. Biol. Chem. 266:2313-2319(1991).

[2] Ostanin K., Harms E.H., Stevis P.E., Kuciel R., Zhou M.-M., van Etten R.L. J. Biol. Chem. 267:22830-22836(1992).

[3] Schneider G., Lindqvist Y., Vihko P. EMBO J. 12:2609-2615(1993).

41. Aconitase family signatures

Aconitase (aconitate hydratase) (EC 4.2.1.3) [1] is the enzyme from the tricarboxylic acid cycle that catalyzes the reversible isomerization of citrate and isocitrate. Cis-aconitate is formed as an intermediary product during the course of the reaction. In eukaryotes two isozymes of aconitase are known to exist: one found in the mitochondrial matrix and the other found in the cytoplasm. Aconitase, in its active form, contains a 4Fe-4S iron-sulfur cluster; three cysteine residues have been shown to be ligands of the 4Fe-4S cluster. It has been shown that the aconitase family also contains the following proteins: - Iron-responsive element binding protein (IRE-BP). IRE-BP is a cytosolic protein that binds to iron-responsive elements (IREs). IREs are stem-loop structures found in the 5'UTR of ferritin, and delta aminolevulinic acid synthase mRNAs, and in the 3'UTR of transferrin receptor mRNA. IRE-BP also express aconitase activity. - 3-isopropylmalate dehydratase (EC 4.2.1.33) (isopropylmalate isomerase), the enzyme that catalyzes the second step in the biosynthesis of leucine. - Homoaconitase (EC 4.2.1.36) (homoaconitate hydratase), an enzyme that participates in the alpha-amino adipate pathway of lysine biosynthesis and that converts cis-homoaconitate into homoisocitric acid. - Escherichia coli protein ybhJ

Consensus pattern: [LIVM SEQ ID NO:4)]-x(2)-[GSACIVM SEQ ID NO:27)]-x-[LIV]-[GTIV SEQ ID NO:28)]-[STP]-C-x(0,1)-T-N- [GSTANI SEQ ID NO:29)]-x(4)-[LIVMA SEQ ID NO:30)] [C binds the iron-sulfur center]

Consensus pattern: G-x(2)-[LIVWPQ SEQ ID NO:31)]-x(3)-[GAC]-C-[GSTAM SEQ ID NO:32)]-[LIMPTA SEQ ID NO:33)]-C-[LIMV SEQ ID NO:34)]- [GA] [The two C's bind the iron-sulfur center]-

[1] Gruer M.J., Artymiuk P.J., Guest J.R. Trends Biochem. Sci. 22:3-6(1997).

5 42. Actins signatures

Actins [1 to 4] are highly conserved contractile proteins that are present in all eukaryotic cells. In vertebrates there are three groups of actin isoforms: alpha, beta and gamma. The alpha actins are found in muscle tissues and are a major constituent of the contractile apparatus. The beta and gamma actins co-exists in most cell types as components of the cytoskeleton and as mediators of internal cell motility. In plants [5] there are many isoforms which are probably involved in a variety of functions such as cytoplasmic streaming, cell shape determination, tip growth, graviperception, cell wall deposition, etc. Actin exists either in a monomeric form (G-actin) or in a polymerized form (F-actin). Each actin monomer can bind a molecule of ATP; when polymerization occurs, the ATP is hydrolyzed. Actin is a protein of from 374 to 379 amino acid residues. The structure of actin has been highly conserved in the course of evolution. Recently some divergent actin-like proteins have been identified in several species. These proteins are: - Centractin (actin-RPV) from mammals, fungi (yeast ACT5, *Neurospora crassa* ro-4) and *Pneumocystis carinii* (actin-II). Centractin seems to be a component of a multi-subunit centrosomal complex involved in microtubule based vesicle motility. This subfamily is also known as ARP1. - ARP2 subfamily which includes chicken ACTL, yeast ACT2, *Drosophila* 14D, *C.elegans* actC. - ARP3 subfamily which includes actin 2 from mammals, *Drosophila* 66B, yeast ACT4 and fission yeast act2. - ARP4 subfamily which includes yeast ACT3 and *Drosophila* 13E. Three signature patterns have been developed. The first two are specific to actins and span positions 54 to 64 and 357 to 365. The last signature picks up both actins and the actin-like proteins and corresponds to positions 106 to 118 in actins.

Consensus pattern: [FY]-[LIV]-G-[DE]-E-A-Q-x-[RKQ](2)-G-

Consensus pattern: W-[IV]-[STA]-[RK]-x-[DE]-Y-[DNE]-[DE]-

30 Consensus pattern: [LM]-[LIVM SEQ ID NO:4])-T-E-[GAPQ SEQ ID NO:56))-x-[LIVMFYWHQ SEQ ID NO:57))-N-[PSTAQ SEQ ID NO:58))-x(2)-N-[KR]-

[1] Sheterline P., Clayton J., Sparrow J.C. (In) Actins, 3rd Edition, Academic Press Ltd, London, (1996).

[2] Pollard T.D., Cooper J.A. Annu. Rev. Biochem. 55:987-1036(1986).

[3] Pollard T.D. Curr. Opin. Cell Biol. 1:33-40(1990).

5 [4] Rubenstein P.A. BioEssays 12:309-315(1990).

[5] Meagher R.B., McLean B.G. Cell Motil. Cytoskeleton 16:164-166(1990).

43. Adenylate kinase signature

10 Adenylate kinase (EC 2.7.4.3) (AK) [1] is a small monomeric enzyme that catalyzes the reversible transfer of MgATP to AMP ($\text{MgATP} + \text{AMP} = \text{MgADP} + \text{ADP}$). In mammals there are three different isozymes: - AK1 (or myokinase), which is cytosolic. - AK2, which is located in the outer compartment of mitochondria. - AK3 (or GTP:AMP phosphotransferase), which is located in the mitochondrial matrix and which uses MgGTP instead of MgATP. The
15 sequence of AK has also been obtained from different bacterial species and from plants and fungi. Two other enzymes have been found to be evolutionary related to AK. These are: - Yeast uridylate kinase (EC 2.7.4.-) (UK) (gene URA6) [2] which catalyzes the transfer of a phosphate group from ATP to UMP to form UDP and ADP. - Slime mold UMP-CMP kinase (EC 2.7.4.14) [3] which catalyzes the transfer of a phosphate group from ATP to either CMP
20 or UMP to form CDP or UDP and ADP. Several regions of AK family enzymes are well conserved, including the ATP-binding domains. The most conserved of all regions have been selected as a signature for this type of enzyme. This region includes an aspartic acid residue that is part of the catalytic cleft of the enzyme and that is involved in a salt bridge. It also includes an arginine residue whose modification leads to inactivation of the enzyme

25

Consensus pattern: [LIVMFYW SEQ ID NO:26)](3)-D-G-[FYI]-P-R-x(3)-[NQ]-

[1] Schulz G.E. Cold Spring Harbor Symp. Quant. Biol. 52:429-439(1987).

30 [2] Liljelund P., Sanni A., Friesen J.D., Lacroute F. Biochem. Biophys. Res. Commun. 165:464-473(1989).

[3] Wiesmueller L., Noegel A.A., Barzu O., Gerisch G., Schleicher M. J. Biol. Chem. 265:6339-6345(1990).

[4] Kath T.H., Schmid R., Schaefer G. Arch. Biochem. Biophys. 307:405-410(1993).

44. (adh_short) Short-chain dehydrogenases/reductases family signature. The short-chain dehydrogenases/reductases family (SDR) [1] is a very large family of enzymes, most of which are known to be NAD- or NADP-dependent oxidoreductases. As the first member of this family to be characterized was *Drosophila* alcohol dehydrogenase, this family used to be called [2,3,4]'insect-type', or 'short-chain' alcohol dehydrogenases. Most member of this family are proteins of about 250 to 300 amino acid residues. The proteins currently known to belong to this family are listed below. - Alcohol dehydrogenase (EC 1.1.1.1) from insects such as *Drosophila*. - Acetoin dehydrogenase (EC 1.1.1.5) from *Klebsiella terrigena* (gene budC). - D-beta-hydroxybutyrate dehydrogenase (BDH) (EC 1.1.1.30) from mammals. - Acetoacetyl-CoA reductase (EC 1.1.1.36) from various bacterial species (gene phbB or phaB). - Glucose 1-dehydrogenase (EC 1.1.1.47) from *Bacillus*. - 3-beta-hydroxysteroid dehydrogenase (EC 1.1.1.51) from *Comomonas testosteroni*. - 20-beta-hydroxysteroid dehydrogenase (EC 1.1.1.53) from *Streptomyces hydrogenans*. - Ribitol dehydrogenase (EC 1.1.1.56) (RDH) from *Klebsiella aerogenes*. - Estradiol 17-beta-dehydrogenase (EC 1.1.1.62) from human. - Gluconate 5-dehydrogenase (EC 1.1.1.69) from *Gluconobacter oxydans* (gene gno). - 3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100) from *Escherichia coli* (gene fabG) and from plants. - Retinol dehydrogenase (EC 1.1.1.105) from mammals. - 2-deoxy-d-gluconate 3-dehydrogenase (EC 1.1.1.125) from *Escherichia coli* and *Erwinia chrysanthemi* (gene kduD). - Sorbitol-6-phosphate 2-dehydrogenase (EC 1.1.1.140) from *Escherichia coli* (gene gutD) and from *Klebsiella pneumoniae* (gene sorD). - 15-hydroxyprostaglandin dehydrogenase (NAD⁺) (EC 1.1.1.141) from human. - Corticosteroid 11-beta-dehydrogenase (EC 1.1.1.146) (11-DH) from mammals. - 7-alpha-hydroxysteroid dehydrogenase (EC 1.1.1.159) from *Escherichia coli* (gene hdhA), *Eubacterium* strain VPI 12708 (gene baiA) and from *Clostridium sordellii*. - NADPH-dependent carbonyl reductase (EC 1.1.1.184) from mammals. - Tropinone reductase-I (EC 1.1.1.206) and -II (EC 1.1.1.236) from plants. - N-acylmannosamine 1-dehydrogenase (EC 1.1.1.233) from *Flavobacterium* strain 141-8. - D-arabinitol 2-dehydrogenase (ribulose forming) (EC 1.1.1.250) from fungi. - Tetrahydroxynaphthalene reductase (EC 1.1.1.252) from *Magnaporthe grisea*. - Pteridine reductase 1 (EC 1.1.1.253) (gene PTR1) from *Leishmania*. - 2,5-dichloro-2,5-cyclohexadiene-1,4-diol dehydrogenase (EC 1.1.-.-) from *Pseudomonas paucimobilis*. - Cis-1,2-dihydroxy-3,4-cyclohexadiene-1-carboxylate dehydrogenase (EC 1.3.1. -) from

Acinetobacter calcoaceticus (gene benD) and Pseudomonas putida (gene xylL). - Biphenyl-2,3-dihydro-2,3-diol dehydrogenase (EC 1.3.1.-) (gene bphB) from various Pseudomonaceae. - Cis-toluene dihydrodiol dehydrogenase (EC 1.3.1.-) from Pseudomonas putida (gene todD). - Cis-benzene glycol dehydrogenase (EC 1.3.1.19) from Pseudomonas putida (gene bnzE). - 2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase (EC 1.3.1.28) from Escherichia coli (gene entA) and Bacillus subtilis (gene dhbA). - Dihydropteridine reductase (EC 1.6.99.7) (HDHPR) from mammals. - Lignin degradation enzyme ligD from Pseudomonas paucimobilis. - Agropine synthesis reductase from Agrobacterium plasmids (gene mas1). - Versicolorin reductase from Aspergillus parasiticus (gene VER1). - Putative keto-acyl reductases from Streptomyces polyketide biosynthesis operons. - A trifunctional hydratase-dehydrogenase-epimerase from the peroxisomal beta-oxidation system of Candida tropicalis. This protein contains two tandemly repeated 'short-chain dehydrogenase-type' domain in its N-terminal extremity. - Nodulation protein nodG from species of Azospirillum and Rhizobium which is probably involved in the modification of the nodulation Nod factor fatty acyl chain. - Nitrogen fixation protein fixR from Bradyrhizobium japonicum. - Bacillus subtilis protein dltE which is involved in the biosynthesis of D- alanyl-lipoteichoic acid. - Human follicular variant translocation protein 1 (FVT1). - Mouse adipocyte protein p27. - Mouse protein Ke 6. - Maize sex determination protein TASSELSEED 2. - Sarcophaga peregrina 25 Kd development specific protein. - Drosophila fat body protein P6. - A Listeria monocytogenes hypothetical protein encoded in the internalins gene region. - Escherichia coli hypothetical protein yciK. - Escherichia coli hypothetical protein ydfG. - Escherichia coli hypothetical protein yjgI. - Escherichia coli hypothetical protein yjgU. - Escherichia coli hypothetical protein yohF. - Bacillus subtilis hypothetical protein yoxD. - Bacillus subtilis hypothetical protein ywfD. - Bacillus subtilis hypothetical protein ywfH. - Yeast hypothetical protein YIL124w. - Yeast hypothetical protein YIR035c. - Yeast hypothetical protein YIR036c. - Yeast hypothetical protein YKL055c. - Fission yeast hypothetical protein SpAC23D3.11. One of the best conserved regions which includes two perfectly conserved residues, a tyrosine and a lysine has been selected as a signature pattern for this family of proteins. The tyrosine residue participates in the catalytic mechanism.

Consensus pattern: [LIVSPADNK SEQ ID NO:59]-x(12)-Y-[PSTAGNCV SEQ ID NO:60)]-[STAGNQCIVM SEQ ID NO:61)]-[STAGC SEQ ID NO:45)]-K- {PC}-[SAGFYR SEQ ID NO:62)]-[LIVMSTAGD SEQ ID NO:63)]-x(2)-[LIVMFYW SEQ ID NO:26)]-x(3)-

[LIVMFYWGAPTHQ SEQ ID NO:64)]-[GSACQRHM SEQ ID NO:65)] [Y is an active site residue] -

[1] Joernvall H., Persson B., Krook M., Atrian S., Gonzalez-Duarte R., Jeffery J., Ghosh D.
5 Biochemistry 34:6003-6013(1995).

[2] Villarroya A., Juan E., Egestad B., Joernvall H. Eur. J. Biochem. 180:191-197(1989).

[3] Persson B., Krook M., Joernvall H. Eur. J. Biochem. 200:537-543(1991).

[4] Neidle E.L., Hartnett C., Ornston N.L., Bairoch A., Rekik M., Harayama S. Eur. J.
10 Biochem. 204:113-120(1992).

45. (adh_short_C2) Short-chain dehydrogenases/reductases family signature

The short-chain dehydrogenases/reductases family (SDR) [1] is a very large family of
enzymes, most of which are known to be NAD- or NADP-dependent oxidoreductases. As the
15 first member of this family to be characterized was *Drosophila* alcohol dehydrogenase, this
family used to be called [2,3,4]'insect-type', or 'short-chain' alcohol dehydrogenases. Most
member of this family are proteins of about 250 to 300 amino acid residues. The proteins
currently known to belong to this family are listed below. - Alcohol dehydrogenase (EC
1.1.1.1) from insects such as *Drosophila*. - Acetoin dehydrogenase (EC 1.1.1.5) from
20 *Klebsiella terrigena* (gene budC). - D-beta-hydroxybutyrate dehydrogenase (BDH) (EC
1.1.1.30) from mammals. - Acetoacetyl-CoA reductase (EC 1.1.1.36) from various bacterial
species (gene phbB or phaB). - Glucose 1-dehydrogenase (EC 1.1.1.47) from *Bacillus*. - 3-
beta-hydroxysteroid dehydrogenase (EC 1.1.1.51) from *Comomonas testosteroni*. - 20-beta-
hydroxysteroid dehydrogenase (EC 1.1.1.53) from *Streptomyces hydrogenans*. - Ribitol
25 dehydrogenase (EC 1.1.1.56) (RDH) from *Klebsiella aerogenes*. - Estradiol 17-beta-
dehydrogenase (EC 1.1.1.62) from human. - Gluconate 5-dehydrogenase (EC 1.1.1.69) from
Gluconobacter oxydans (gene gno). - 3-oxoacyl-[acyl-carrier protein] reductase (EC
1.1.1.100) from *Escherichia coli* (gene fabG) and from plants. - Retinol dehydrogenase (EC
1.1.1.105) from mammals. - 2-deoxy-d-gluconate 3-dehydrogenase (EC 1.1.1.125) from
30 *Escherichia coli* and *Erwinia chrysanthemi* (gene kduD). - Sorbitol-6-phosphate 2-
dehydrogenase (EC 1.1.1.140) from *Escherichia coli* (gene gutD) and from *Klebsiella*
pneumoniae (gene sorD). - 15-hydroxyprostaglandin dehydrogenase (NAD⁺) (EC 1.1.1.141)
from human. - Corticosteroid 11-beta-dehydrogenase (EC 1.1.1.146) (11-DH) from

- mammals. - 7-alpha-hydroxysteroid dehydrogenase (EC 1.1.1.159) from *Escherichia coli* (gene *hdhA*), *Eubacterium* strain VPI 12708 (gene *baiA*) and from *Clostridium sordellii*. - NADPH-dependent carbonyl reductase (EC 1.1.1.184) from mammals. - Tropinone reductase-I (EC 1.1.1.206) and -II (EC 1.1.1.236) from plants. - N-acylmannosamine 1-
- 5 dehydrogenase (EC 1.1.1.233) from *Flavobacterium* strain 141-8. - D-arabinitol 2-dehydrogenase (ribulose forming) (EC 1.1.1.250) from fungi. - Tetrahydroxynaphthalene reductase (EC 1.1.1.252) from *Magnaporthe grisea*. - Pteridine reductase 1 (EC 1.1.1.253) (gene *PTR1*) from *Leishmania*. - 2,5-dichloro-2,5-cyclohexadiene-1,4-diol dehydrogenase (EC 1.1.-.-) from *Pseudomonas paucimobilis*. - Cis-1,2-dihydroxy-3,4-cyclohexadiene-1-
- 10 carboxylate dehydrogenase (EC 1.3.1. -) from *Acinetobacter calcoaceticus* (gene *benD*) and *Pseudomonas putida* (gene *xylL*). - Biphenyl-2,3-dihydro-2,3-diol dehydrogenase (EC 1.3.1.-) (gene *bphB*) from various *Pseudomonaceae*. - Cis-toluene dihydrodiol dehydrogenase (EC 1.3.1.-) from *Pseudomonas putida* (gene *todD*). - Cis-benzene glycol dehydrogenase (EC 1.3.1.19) from *Pseudomonas putida* (gene *bnzE*). - 2,3-dihydro-2,3-dihydroxybenzoate
- 15 dehydrogenase (EC 1.3.1.28) from *Escherichia coli* (gene *entA*) and *Bacillus subtilis* (gene *dhbA*). - Dihydropteridine reductase (EC 1.6.99.7) (HDHPR) from mammals. - Lignin degradation enzyme *ligD* from *Pseudomonas paucimobilis*. - Agropine synthesis reductase from *Agrobacterium* plasmids (gene *mas1*). - Versicolorin reductase from *Aspergillus parasiticus* (gene *VER1*). - Putative keto-acyl reductases from *Streptomyces* polyketide
- 20 biosynthesis operons. - A trifunctional hydratase-dehydrogenase-epimerase from the peroxisomal beta-oxidation system of *Candida tropicalis*. This protein contains two tandemly repeated 'short-chain dehydrogenase-type' domain in its N-terminal extremity. - Nodulation protein *nodG* from species of *Azospirillum* and *Rhizobium* which is probably involved in the modification of the nodulation Nod factor fatty acyl chain. - Nitrogen fixation protein *fixR*
- 25 from *Bradyrhizobium japonicum*. - *Bacillus subtilis* protein *dltE* which is involved in the biosynthesis of D- alanyl-lipoteichoic acid. - Human follicular variant translocation protein 1 (FVT1). - Mouse adipocyte protein p27. - Mouse protein Ke 6. - Maize sex determination protein TASSELSEED 2. - *Sarcophaga peregrina* 25 Kd development specific protein. - *Drosophila* fat body protein P6. - A *Listeria monocytogenes* hypothetical protein encoded in
- 30 the internalins gene region. - *Escherichia coli* hypothetical protein *yciK*. - *Escherichia coli* hypothetical protein *ydfG*. - *Escherichia coli* hypothetical protein *yjgI*. - *Escherichia coli* hypothetical protein *yjgU*. - *Escherichia coli* hypothetical protein *yohF*. - *Bacillus subtilis* hypothetical protein *yoxD*. - *Bacillus subtilis* hypothetical protein *ywfD*. - *Bacillus subtilis*

hypothetical protein ywfH. - Yeast hypothetical protein YIL124w. - Yeast hypothetical protein YIR035c. - Yeast hypothetical protein YIR036c. - Yeast hypothetical protein YKL055c. - Fission yeast hypothetical protein SpAC23D3.11. One of the best conserved regions which includes two perfectly conserved residues, a tyrosine and a lysine has been used as a signature pattern for this family of proteins. The tyrosine residue participates in the catalytic mechanism.

Consensus pattern: [LIVSPADNK SEQ ID NO:59)]-x(12)-Y-[PSTAGNCV SEQ ID NO:60)]-[STAGNQCIVM SEQ ID NO:61)]-[STAGC SEQ ID NO:45)]-K- {PC}-[SAGFYR SEQ ID NO:62)]-[LIVMSTAGD SEQ ID NO:63)]-x(2)-[LIVMFYW SEQ ID NO:26)]-x(3)-[LIVMFYWGAPTHQ SEQ ID NO:64)]-[GSACQRHM SEQ ID NO:65)] [Y is an active site residue]

[1] Joernvall H., Persson B., Krook M., Atrian S., Gonzalez-Duarte R., Jeffery J., Ghosh D. Biochemistry 34:6003-6013(1995).

[2] Villarroya A., Juan E., Egestad B., Joernvall H. Eur. J. Biochem. 180:191-197(1989).

[3] Persson B., Krook M., Joernvall H. Eur. J. Biochem. 200:537-543(1991).

[4] Neidle E.L., Hartnett C., Ornston N.L., Bairoch A., Rekik M., Harayama S. Eur. J. Biochem. 204:113-120(1992).

46. (adh_zinc) Zinc-containing alcohol dehydrogenases signatures

Alcohol dehydrogenase (EC 1.1.1.1) (ADH) catalyzes the reversible oxidation of ethanol to acetaldehyde with the concomitant reduction of NAD [1]. Currently three, structurally and catalytically, different types of alcohol dehydrogenases are known: - Zinc-containing 'long-chain' alcohol dehydrogenases. - Insect-type, or 'short-chain' alcohol dehydrogenases. - Iron-containing alcohol dehydrogenases. Zinc-containing ADH's [2,3] are dimeric or tetrameric enzymes that bind two atoms of zinc per subunit. One of the zinc atom is essential for catalytic activity while the other is not. Both zinc atoms are coordinated by either cysteine or histidine residues; the catalytic zinc is coordinated by two cysteines and one histidine. Zinc-containing ADH's are found in bacteria, mammals, plants, and in fungi. In most species there are more than one isozyme (for example, human have at least six isozymes, yeast have three, etc.). A number of other zinc-dependent dehydrogenases are closely related to zinc ADH [4],

- these are: - Xylitol dehydrogenase (EC 1.1.1.9) (D-xylulose reductase). - Sorbitol dehydrogenase (EC 1.1.1.14). - Aryl-alcohol dehydrogenase (EC 1.1.1.90) (benzyl alcohol dehydrogenase). - Threonine 3-dehydrogenase (EC 1.1.1.103). - Cinnamyl-alcohol dehydrogenase (EC 1.1.1.195) (CAD) [5]. CAD is a plant enzyme involved in the biosynthesis of lignin. - Galactitol-1-phosphate dehydrogenase (EC 1.1.1.251). - *Pseudomonas putida* 5-exo-alcohol dehydrogenase (EC 1.1.1.-) [6]. - *Escherichia coli* starvation sensing protein *rspB*. - *Escherichia coli* hypothetical protein *yjgB*. - *Escherichia coli* hypothetical protein *yjgV*. - *Escherichia coli* hypothetical protein *yjiN*. - Yeast hypothetical protein YAL060w (FUN49). - Yeast hypothetical protein YAL061w (FUN50). - Yeast hypothetical protein YCR105w. The pattern that has been developed to detect this class of enzymes is based on a conserved region that includes a histidine residue which is the second ligand of the catalytic zinc atom. This family also includes NADP-dependent quinone oxidoreductase (EC 1.6.5.5), an enzyme found in bacteria (gene *qor*), in yeast and in mammals where, in some species such as rodents, it has been recruited as an eye lens protein and is known as zeta-crystallin [7]. The sequence of quinone oxidoreductase is distantly related to that of other zinc-containing alcohol dehydrogenases and it lacks the zinc-ligand residues. The torpedo fish and mammalian synaptic vesicle membrane protein *vat-1* is related to *qor*. A specific pattern has been developed for this subfamily.
- Consensus pattern: G-H-E-x(2)-G-x(5)-[GA]-x(2)-[IVSAC SEQ ID NO:66] [H is a zinc ligand]
 Consensus pattern: [GSD]-[DEQH SEQ ID NO:67]-x(2)-L-x(3)-[SA](2)-G-G-x-G-x(4)-Q-x(2)-[KR]-
- [1] Branden C.-I., Joernvall H., Eklund H., Furugren B. (In) The Enzymes (3rd edition) 11:104-190(1975).
 [2] Joernvall H., Persson B., Jeffery J. Eur. J. Biochem. 167:195-201(1987).
 [3] Sun H.-W., Plapp B.V. J. Mol. Evol. 34:522-535(1992).
 [4] Persson B., Hallborn J., Walfridsson M., Hahn-Haegerdal B., Keraenen S., Penttilae M., Joernvall H. FEBS Lett. 324:9-14(1993).
 [5] Knight M.E., Halpin C., Schuch W. Plant Mol. Biol. 19:793-801(1992).
 [6] Koga H., Aramaki H., Yamaguchi E., Takeuchi K., Horiuchi T., Gunsalus I.C. J. Bacteriol. 166:1089-1095(1986).

[7] Joernvall H., Persson B., Du Bois G., Lavers G.C., Chen J.H., Gonzalez P., Rao P.V., Zigler J.S. Jr. FEBS Lett. 322:240-244(1993).

- 5 47. (aldedh) Aldehyde dehydrogenases active sites
- Aldehyde dehydrogenases (EC 1.2.1.3 and EC 1.2.1.5) are enzymes which oxidize a wide variety of aliphatic and aromatic aldehydes. In mammals at least four different forms of the enzyme are known [1]: class-1 (or Ald C) a tetrameric cytosolic enzyme, class-2 (or Ald M) a tetrameric mitochondrial enzyme, class-3 (or Ald D) a dimeric cytosolic enzyme, and class
- 10 IV a microsomal enzyme. Aldehyde dehydrogenases have also been sequenced from fungal and bacterial species. A number of enzymes are known to be evolutionary related to aldehyde dehydrogenases; these enzymes are listed below. - Plants and bacterial betaine-aldehyde dehydrogenase (EC 1.2.1.8) [2], an enzyme that catalyzes the last step in the biosynthesis of betaine. - Plants and bacterial NADP-dependent glyceraldehyde-3-phosphate dehydrogenase
- 15 (EC 1.2.1.9). - Escherichia coli succinate-semialdehyde dehydrogenase (NADP+) (EC 1.2.1.16) (gene gabD) [3], which reduces succinate semialdehyde into succinate. - Escherichia coli lactaldehyde dehydrogenase (EC 1.2.1.22) (gene ald) [4]. - Mammalian succinate semialdehyde dehydrogenase (NAD+) (EC 1.2.1.24). - Escherichia coli phenylacetaldehyde dehydrogenase (EC 1.2.1.39). - Escherichia coli 5-carboxymethyl-2-
- 20 hydroxymuconate semialdehyde dehydrogenase (gene hpcC). - Pseudomonas putida 2-hydroxymuconic semialdehyde dehydrogenase [5] (genes dmpC and xylG), an enzyme in the meta-cleavage pathway for the degradation of phenols, cresols and catechol. - Bacterial and mammalian methylmalonate-semialdehyde dehydrogenase (MMSDH) (EC 1.2.1.27) [6], an enzyme involved in the distal pathway of valine catabolism. - Yeast delta-1-pyrroline-5-
- 25 carboxylate dehydrogenase (EC 1.5.1.12) [7] (gene PUT2), which converts proline to glutamate. - Bacterial multifunctional putA protein, which contains a delta-1-pyrroline- 5-carboxylate dehydrogenase domain. - 26G, a garden pea protein of unknown function which is induced by dehydration of shoots [8]. - Mammalian formyltetrahydrofolate dehydrogenase (EC 1.5.1.6) [9]. This is a cytosolic enzyme responsible for the NADP-dependent
- 30 decarboxylative reduction of 10-formyltetrahydrofolate into tetrahydrofolate. It is a protein of about 900 amino acids which consist of three domains; the C- terminal domain (480 residues) is structurally and functionally related to aldehyde dehydrogenases. - Yeast hypothetical protein YBR006w. - Yeast hypothetical protein YER073w. - Yeast hypothetical

protein YHR039c. - *Caenorhabditis elegans* hypothetical protein F01F1.6.A glutamic acid and a cysteine residue have been implicated in the catalytic activity of mammalian aldehyde dehydrogenase. These residues are conserved in all the enzymes of this family. Two patterns have been derived for this family, one for each of the active site residues.

5

Consensus pattern: [LIVMFGA SEQ ID NO:68]-E-[LIMSTAC SEQ ID NO:69)]-[GS]-G-[KNLM SEQ ID NO:70)]-[SADN SEQ ID NO:71)]-[TAPFV SEQ ID NO:72)] [E is the active site residue]-

10

Consensus pattern: [FYLVA SEQ ID NO:73)]-x(3)-G-[QE]-x-C-[LIVMGSTANC SEQ ID NO:74)]-[AGCN SEQ ID NO:75)]-x-[GSTADNEKR SEQ ID NO:76)] [C is the active site residue]

[1] Hempel J., Harper K., Lindahl R. Biochemistry 28:1160-1167(1989).

[2] Weretilnyk E.A., Hanson A.D. Proc. Natl. Acad. Sci. U.S.A. 87:2745-2749(1990).

15

[3] Niegemann E., Schulz A., Bartsch K. Arch. Microbiol. 160:454-460(1993).

[4] Hidalgo E., Chen Y.-M., Lin E.C.C., Aguilar J. J. Bacteriol. 173:6118-6123(1991).

[5] Nordlund I., Shingler V. Biochim. Biophys. Acta 1049:227-230(1990).

[6] Steele M.I., Lorenz D., Hatter K., Park A., Sokatch J.R. J. Biol. Chem. 267:13585-13592(1992).

20

[7] Krzywicki K.A., Brandriss M.C. Mol. Cell. Biol. 4:2837-2842(1984).

[8] Guerrero F.D., Jones J.T., Mullet J.E. Plant Mol. Biol. 15:11-26(1990).

[9] Cook R.J., Lloyd R.S., Wagner C. J. Biol. Chem. 266:4965-4973(1991).

25

48. Aldo/keto reductase family signatures

The aldo-keto reductase family [1,2] groups together a number of structurally and functionally related NADPH-dependent oxidoreductases as well as some other proteins. The proteins known to belong to this family are: - Aldehyde reductase (EC 1.1.1.2). - Aldose reductase (EC 1.1.1.21). - 3-alpha-hydroxysteroid dehydrogenase (EC 1.1.1.50), which terminates androgen action by converting 5-alpha-dihydrotestosterone to 3-alpha-androstanediol. - Prostaglandin F synthase (EC 1.1.1.188) which catalyzes the reduction of prostaglandins H2 and D2 to F2-alpha. - D-sorbitol-6-phosphate dehydrogenase (EC 1.1.1.200) from apple. - Morphine 6-dehydrogenase (EC 1.1.1.218) from *Pseudomonas*

30

putida plasmid pMDH7.2 (gene *morA*). - Chlordecone reductase (EC 1.1.1.225) which reduces the pesticide chlordecone (kepone) to the corresponding alcohol. - 2,5-diketo-D-gluconic acid reductase (EC 1.1.1.-) which catalyzes the reduction of 2,5-diketogluconic acid to 2-keto-L-gulonic acid, a key intermediate in the production of ascorbic acid. - NAD(P)H-dependent xylose reductase (EC 1.1.1.-) from the yeast *Pichia stipitis*. This enzyme reduces xylose into xylitol. - Trans-1,2-dihydrobenzene-1,2-diol dehydrogenase (EC 1.3.1.20). - 3-oxo-5-beta-steroid 4-dehydrogenase (EC 1.3.99.6) which catalyzes the reduction of delta(4)-3-oxosteroids. - A soybean reductase, which co-acts with chalcone synthase in the formation of 4,2',4'-trihydroxychalcone. - Frog eye lens rho crystallin. - Yeast GCY protein, whose function is not known. - *Leishmania major* P110/11E protein. P110/11E is a developmentally regulated protein whose abundance is markedly elevated in promastigotes compared with amastigotes. Its exact function is not yet known. - *Escherichia coli* hypothetical protein *yafB*. - *Escherichia coli* hypothetical protein *yghE*. - Yeast hypothetical protein YBR149w. - Yeast hypothetical protein YHR104w. - Yeast hypothetical protein YJR096w. These proteins have all about 300 amino acid residues. Three consensus patterns have been developed that are specific to this family of proteins. The first one is located in the N-terminal section of these proteins. The second pattern is located in the central section. The third pattern, located in the C-terminal, is centered on a lysine residue whose chemical modification, in aldose and aldehydereductases, affect the catalytic efficiency.

Consensus pattern: G-[FY]-R-[HSAL SEQ ID NO:77)]-[LIVMF SEQ ID NO:2)]-D-[STAGC SEQ ID NO:45)]-[AS]-x(5)-E-x(2)-[LIVM SEQ ID NO:4)]- G -

Consensus pattern: [LIVMFY SEQ ID NO:18)]-x(9)-[KREQ SEQ ID NO:78)]-x-[LIVM SEQ ID NO:4)]-G-[LIVM SEQ ID NO:4)]-[SC]-N-[FY]-

Consensus pattern: [LIVM SEQ ID NO:4)]-[PAIV SEQ ID NO:79)]-[KR]-[ST]-x(4)-R-x(2)-[GSTAEQK SEQ ID NO:80)]-[NSL]-x(2)- [LIVMFA SEQ ID NO:81)] [K is a putative active site residue]-

[1] Bohren K.M., Bullock B., Wermuth B., Gabbay K.H. J. Biol. Chem. 264:9547-9551(1989).

[2] Bruce N.C., Willey D.L., Coulson A.F.W., Jeffery J. Biochem. J. 299:805-811(1994).

49. Alpha amylase. This family is classified as family 13 of the glycosyl hydrolases. The structure is an 8 stranded alpha/beta barrel, interrupted by a ~70 a.a. calcium-binding domain protruding between beta strand 3 and alpha helix 3, and a carboxyl-terminal Greek key beta-barrel domain.

5

[1] Larson SB, Greenwood A, Cascio D, Day J, McPherson A, J Mol Biol 1994;235:1560-1584.

10 50. Aminotransferases class-I pyridoxal-phosphate attachment site

Aminotransferases share certain mechanistic features with other pyridoxal- phosphate dependent enzymes, such as the covalent binding of the pyridoxal- phosphate group to a lysine residue. On the basis of sequence similarity, these various enzymes can be grouped [1,2] into subfamilies. One of these, called class-I, currently consists of the following
 15 enzymes: - Aspartate aminotransferase (AAT) (EC 2.6.1.1). AAT catalyzes the reversible transfer of the amino group from L-aspartate to 2-oxoglutarate to form oxaloacetate and L-glutamate. In eukaryotes, there are two AAT isozymes: one is located in the mitochondrial matrix, the second is cytoplasmic. In prokaryotes, only one form of AAT is found (gene aspC). - Tyrosine aminotransferase (EC 2.6.1.5) which catalyzes the first step in tyrosine
 20 catabolism by reversibly transferring its amino group to 2- oxoglutarate to form 4-hydroxyphenylpyruvate and L-glutamate. - Aromatic aminotransferase (EC 2.6.1.57) involved in the synthesis of Phe, Tyr, Asp and Leu (gene tyrB). - 1-aminocyclopropane-1-carboxylate synthase (EC 4.4.1.14) (ACC synthase) from plants. ACC synthase catalyzes the first step in ethylene biosynthesis. - Pseudomonas denitrificans cobC, which is involved in
 25 cobalamin biosynthesis. - Yeast hypothetical protein YJL060w. The sequence around the pyridoxal-phosphate attachment site of this class of enzyme is sufficiently conserved to allow the creation of a specific pattern.

Consensus pattern: [GS]-[LIVMFYTAC SEQ ID NO:82)]-[GSTA SEQ ID NO:19)]-K-x(2)-
 30 [GSALVN SEQ ID NO:83)]-[LIVMFA SEQ ID NO:81)]-x-[GNAR SEQ ID NO:84)]- x-R-[LIVMA SEQ ID NO:30)]-[GA] [K is the pyridoxal-P attachment site]

[1] Bairoch A. Unpublished observations (1992).

[2] Sung M.H., Tanizawa K., Tanaka H., Kuramitsu S., Kagamiyama H., Hirotsu K., Okamoto A., Higuchi T., Soda K. J. Biol. Chem. 266:2567-2572(1991).

5 51. Aminotransferases class-II pyridoxal-phosphate attachment site

Aminotransferases share certain mechanistic features with other pyridoxal- phosphate dependent enzymes, such as the covalent binding of the pyridoxal- phosphate group to a lysine residue. On the basis of sequence similarity, these various enzymes can be grouped [1] into subfamilies. One of these, called class-II, currently consists of the following enzymes: -

- 10 Glycine acetyltransferase (EC 2.3.1.29), which catalyzes the addition of acetyl-CoA to glycine to form 2-amino-3-oxobutanoate (gene kbl). - 5-aminolevulinic acid synthase (EC 2.3.1.37) (delta-ALA synthase), which catalyzes the first step in heme biosynthesis via the Shemin (or C4) pathway, i.e. the addition of succinyl-CoA to glycine to form 5-aminolevulinate. - 8-amino-7-oxononanoate synthase (EC 2.3.1.47) (7-KAP synthetase), a
15 bacterial enzyme (gene bioF) which catalyzes an intermediate step in the biosynthesis of biotin: the addition of 6-carboxy-hexanoyl-CoA to alanine to form 8-amino-7-oxononanoate. - Histidinol-phosphate aminotransferase (EC 2.6.1.9), which catalyzes the eighth step in histidine biosynthetic pathway: the transfer of an amino group from 3-(imidazol-4-yl)-2-oxopropyl phosphate to glutamic acid to form histidinol phosphate and 2-oxoglutarate. -
20 Serine palmitoyltransferase (EC 2.3.1.50) from yeast (genes LCB1 and LCB2), which catalyzes the condensation of palmitoyl-CoA and serine to form 3- ketosphinganine. The sequence around the pyridoxal-phosphate attachment site of this class of enzyme is sufficiently conserved to allow the creation of a specific pattern

25 Consensus pattern: T-[LIVMFYW SEQ ID NO:26)]-[STAG SEQ ID NO:20)]-K-[SAG]-[LIVMFYWR SEQ ID NO:85)]-[SAG]-x(2)-[SAG] [K is the pyridoxal-P attachment site]-

[1] Bairoch A. Unpublished observations (1991).

30

52. Aminotransferases class-III pyridoxal-phosphate attachment site

Aminotransferases share certain mechanistic features with other pyridoxal- phosphate dependent enzymes, such as the covalent binding of the pyridoxal- phosphate group to a

lysine residue. On the basis of sequence similarity, these various enzymes can be grouped [1,2] into subfamilies. One of these, called class-III, currently consists of the following enzymes: - Acetylornithine aminotransferase (EC 2.6.1.11) which catalyzes the transfer of an amino group from acetylornithine to alpha-ketoglutarate, yielding N-acetyl-glutamic-5-semi-

5 aldehyde and glutamic acid. - Ornithine aminotransferase (EC 2.6.1.13), which catalyzes the transfer of an amino group from ornithine to alpha-ketoglutarate, yielding glutamic-5- semi-

aldehyde and glutamic acid. - Omega-amino acid--pyruvate aminotransferase (EC 2.6.1.18), which catalyzes transamination between a variety of omega-amino acids, mono- and

diamines, and pyruvate. It plays a pivotal role in omega amino acids metabolism. - 4-

10 aminobutyrate aminotransferase (EC 2.6.1.19) (GABA transaminase), which catalyzes the transfer of an amino group from GABA to alpha-ketoglutarate, yielding succinate

semialdehyde and glutamic acid. - DAPA aminotransferase (EC 2.6.1.62), a bacterial enzyme (gene bioA) which catalyzes an intermediate step in the biosynthesis of biotin, the

transamination of 7-keto-8-aminopelargonic acid (7-KAP) to form 7,8- diaminopelargonic

15 acid (DAPA). - 2,2-dialkylglycine decarboxylase (EC 4.1.1.64), a *Pseudomonas cepacia* enzyme (gene dgdA) that catalyzes the decarboxylating amino transfer of 2,2-dialkylglycine

and pyruvate to dialkyl ketone, alanine and carbon dioxide. - Glutamate-1-semialdehyde

aminotransferase (EC 5.4.3.8) (GSA). GSA is the enzyme involved in the second step of

porphyrin biosynthesis, via the C5 pathway. It transfers the amino group on carbon 2 of

20 glutamate-1- semialdehyde to the neighbouring carbon, to give delta-aminolevulinic acid. -

Bacillus subtilis aminotransferase yhxA. - *Bacillus subtilis* aminotransferase yodT. -

Haemophilus influenzae aminotransferase HI0949. - *Caenorhabditis elegans* aminotransferase T01B11.2. The sequence around the pyridoxal-phosphate attachment site of this class

of enzyme is sufficiently conserved to allow the creation of a specific pattern.

25 Consensus pattern: [LIVMFYWC SEQ ID NO:86]](2)-x-D-E-[IVA]-x(2)-G-[LIVMFAGC SEQ ID NO:87)]-x(0,1)- [RSACLI SEQ ID NO:88)]-x-[GSAD SEQ ID NO:89)]-x(12,16)-D-[LIVMFC SEQ ID NO:90)]-[LIVMFYSTA SEQ ID NO:91)]-x(2)- [GSA]-K-x(3)-[GSTADNV SEQ ID NO:92)]-[GSAC SEQ ID NO:93)] [K is the pyridoxal-P attachment

30 site]-

[1] Bairoch A. Unpublished observations (1992).[2] Yonaha K., Nishie M., Aibara S. J. Biol. Chem. 267:12506-12510(1992).

53. Ank repeat. There's no clear separation between noise and signal on the HMM search
Ankyrin repeats generally consist of a beta, alpha, alpha, beta order of secondary structures.
5 The repeats associate to form a higher order structure.

[1] A, Holak TA, FEBS Lett 1997;401:127-132.

[2] Lux SE, John KM, Bennett V, Nature 1990;345:736-739.

10 54. Aminotransferases class-IV signature

Aminotransferases share certain mechanistic features with other pyridoxal-phosphate
dependent enzymes, such as the covalent binding of the pyridoxal-phosphate group to a
lysine residue. On the basis of sequence similarity, these various enzymes can be grouped
15 [1,2] into subfamilies. One of these, called class-IV, currently consists of the following
enzymes:

- Branched-chain amino-acid aminotransferase (EC 2.6.1.42) (transaminase B), a
bacterial (gene *ilvE*) and eukaryotic enzyme which catalyzes the reversible
transfer of an amino group from 4-methyl-2-oxopentanoate to glutamate, to form
20 leucine and 2-oxoglutarate.
- D-alanine aminotransferase (EC 2.6.1.21). A bacterial enzyme which catalyzes the
transfer of the amino group from D-alanine (and other D-amino acids) to 2-
oxoglutarate, to form pyruvate and D-aspartate.
- 4-amino-4-deoxychorismate (ADC) lyase (gene *pabC*). A bacterial enzyme that
25 converts ADC into 4-aminobenzoate (PABA) and pyruvate.

The above enzymes are proteins of about 270 to 415 amino-acid residues that share a
few regions of sequence similarity. Surprisingly, the best-conserved region does not include
the lysine residue to which the pyridoxal-phosphate group is known to be attached, in *ilvE*.
The region that has been selected as a signature pattern is located some 40 residues at the C-
30 terminus side of the PIP-lysine

Consensus pattern: E-x-[STAGCI SEQ ID NO:94)]-x(2)-N-[LIVMFAC SEQ ID NO:95)]-[FY]-x(6,12)-[LIVMF SEQ ID NO:2)]-x-T-x(6,8)-[LIVM SEQ ID NO:4)]-x-[GS]-[LIVM SEQ ID NO:4)]-x-[KR]-

- 5 [1] Green J.M., Merkel W.K., Nichols B.P. J. Bacteriol. 174:5317-5323(1992).
[2] Bairoch A. Unpublished observations (1992).

55. Aminotransferases class-V pyridoxal-phosphate attachment site

Aminotransferases share certain mechanistic features with other pyridoxal- phosphate
10 dependent enzymes, such as the covalent binding of the pyridoxal- phosphate group to a
lysine residue. On the basis of sequence similarity, these various enzymes can be grouped
[1,2] into subfamilies. One of these, called class-V, currently consists of the following
enzymes: - Phosphoserine aminotransferase (EC 2.6.1.52), an enzyme which catalyzes the
reversible interconversion of phosphoserine and 2-oxoglutarate to 3-phosphonooxypyruvate
15 and glutamate. It is required both in the major phosphorylated pathway of serine biosynthesis
and in pyridoxine biosynthesis. The bacterial enzyme (gene serC) is highly similar to a rabbit
endometrial progesterone-induced protein (EPIP), which is probably a phosphoserine
aminotransferase [3]. - Serine--glyoxylate aminotransferase (EC 2.6.1.45) (SGAT) (gene
sgaA) from *Methylobacterium extorquens*. - Serine--pyruvate aminotransferase (EC
20 2.6.1.51). This enzyme also acts as an alanine--glyoxylate aminotransferase (EC 2.6.1.44). In
vertebrates, it is located in the peroxisomes and/or mitochondria. - Isopenicillin N epimerase
(gene cefD). This enzyme is involved in the biosynthesis of cephalosporin antibiotics and
catalyzes the reversible isomerization of isopenicillin N and penicillin N. - NifS, a protein of
the nitrogen fixation operon of some bacteria and cyanobacteria. The exact function of nifS is
25 not yet known. A highly similar protein has been found in fungi (gene NFS1 or SPL1). - The
small subunit of cyanobacterial soluble hydrogenase (EC 1.12.-.-). - Hypothetical protein
ycbU from *Bacillus subtilis*. - Hypothetical protein YFL030w from yeast. The sequence
around the pyridoxal-phosphate attachment site of this class of enzyme is sufficiently
conserved to allow the creation of a specific pattern.

30

Consensus pattern: [LIVFYCHT SEQ ID NO:96)]-[DGH]-[LIVMFYAC SEQ ID NO:97)]-[LIVMFYA SEQ ID NO:98)]-x(2)-[GSTAC SEQ ID NO:99)]-[GSTA SEQ ID NO:19)]-

[HQR]-K-x(4,6)-G-x-[GSAT SEQ ID NO:100)]-x-[LIVMFYSAC SEQ ID NO:101)] [K is the pyridoxal-P attachment site]-

[1] Ouzounis C., Sander C. FEBS Lett. 322:159-164(1993).

5 [2] Bairoch A. Unpublished observations (1992).

[3] van der Zel A., Lam H.-M., Winkler M.E. Nucleic Acids Res. 17:8379-8379(1989).

56. Annexins repeated domain signature

10 Annexins [1 to 6] are a group of calcium-binding proteins that associate reversibly with membranes. They bind to phospholipid bilayers in the presence of micromolar free calcium concentration. The binding is specific for calcium and for acidic phospholipids. Annexins have been claimed to be involved in cytoskeletal interactions, phospholipase inhibition, intracellular signalling, anticoagulation, and membrane fusion. Each of these proteins consist
15 of an N-terminal domain of variable length followed by four or eight copies of a conserved segment of sixty one residues. The repeat (sometimes known as an 'endonexin fold') consists of five alpha-helices that are wound into a right-handed superhelix [7]. The proteins known to belong to the annexin family are listed below: - Annexin I (Lipocortin 1) (Calpactin 2) (p35) (Chromobindin 9). - Annexin II (Lipocortin 2) (Calpactin 1) (Protein I) (p36) (Chromobindin
20 8). - Annexin III (Lipocortin 3) (PAP-III). - Annexin IV (Lipocortin 4) (Endonexin I) (Protein II) (Chromobindin 4). - Annexin V (Lipocortin 5) (Endonexin 2) (VAC-alpha) (Anchorin CII) (PAP-I). - Annexin VI (Lipocortin 6) (Protein III) (Chromobindin 20) (p68) (p70). This is the only known annexin that contains 8 (instead of 4) repeats. - Annexin VII (Synexin). - Annexin VIII (Vascular anticoagulant-beta) (VAC-beta). - Annexin IX from Drosophila. -
25 Annexin X from Drosophila. - Annexin XI (Calcyclin-associated annexin) (CAP-50). - Annexin XII from Hydra vulgaris. - Annexin XIII (Intestine-specific annexin) (ISA). The signature pattern for this domain spans positions 9 to 61 of the repeat and includes the only perfectly conserved residue (an arginine in position 22)-

30 Consensus pattern: [TG]-[STV]-x(8)-[LIVMF SEQ ID NO:2)]-x(2)-R-x(3)-[DEQNH SEQ ID NO:102)]-x(7)-[IFY]- x(7)-[LIVMF SEQ ID NO:2)]-x(3)-[LIVMF SEQ ID NO:2)]-x(11)-[LIVMFA SEQ ID NO:81)]-x(2)-[LIVMF SEQ ID NO:2)]-

- [1] Raynal P., Pollard H.B. *Biochim. Biophys. Acta* 1197:63-93(1994).
- [2] Barton G.J., Newman R.H., Freemont P.S., Crumpton M.J. *Eur. J. Biochem.* 198:749-760(1991).
- [3] Burgoyne R.D., Geisow M.J. *Cell Calcium* 10:1-10(1989).
- 5 [4] Haigler H.T., Fitch J.M., Jones J.M., Schlaepfer D.D. *Trends Biochem. Sci.* 14:48-50(1989).
- [5] Klee C.B. *Biochemistry* 27:6645-6653(1988).
- [6] Smith P.D., Moss S.E. *Trends Genet.* 10:241-246(1994).
- [7] Huber R., Roemisch J., Paques E.-P. *EMBO J.* 9:3867-3874(1990).
- 10 [8] Fiedler K., Simons K. *Trends Biochem. Sci.* 20:177-178(1995).

57. (arf_1) ADP-ribosylation factors family signature

ADP-ribosylation factors (ARF) [1,2,3,4] are 20 Kd GTP-binding proteins involved in
 15 protein trafficking. They may modulate vesicle budding and uncoating within the Golgi apparatus. ARF's also act as allosteric activators of cholera toxin ADP-ribosyltransferase activity. They are evolutionary conserved and present in all eukaryotes. At least six forms of ARF are present in mammals and three in budding yeast. The ARF family also includes
 20 proteins highly related to ARF's but which lack the cholera toxin cofactor activity, they are collectively known as ARL's (ARF-like). ARD1 is a 64 Kd mammalian protein of unknown biological function that contains an ARF domain at its C-terminal extremity. Proteins from the ARF family are generally included in the RAS 'superfamily' of small GTP-binding
 25 proteins [5], but they are only slightly related to the other RAS proteins. They also differ from RAS proteins in that they lack cysteine residues at their C-termini and are therefore not subject to prenylation. The ARFs are N-terminally myristoylated (the ARLs have not yet
 been shown to be modified in such a fashion). A conserved region in the C-terminal part of ARF's and ARL's has been selected as a signature pattern.

Consensus pattern: [HRQT SEQ ID NO:103)]-x-[FYWI SEQ ID NO:104)]-x-[LIVM SEQ ID
 30 NO:4)]-x(4)-A-x(2)-G-x(2)-[LIVM SEQ ID NO:4)]-x(2)-[GSA]-[LIVMF SEQ ID NO:2)]-x-[WK]-[LIVM SEQ ID NO:4)]-

Note: proteins belonging to this family also contain a copy of the ATP/GTP- binding motif 'A' (P-loop) (see <[PDOC00017](#)

- [1] Boman A.L., Kahn R.A. Trends Biochem. Sci. 20:147-150(1995).
 [2] Moss J., Vaughan M. Cell. Signal. 4:367-399(1993).
 [3] Moss J., Vaughan M. Prog. Nucleic Acid Res. Mol. Biol. 45:47-65(1993).
 5 [4] Amor J.C., Harrison D.H., Kahn R.A., Ringe D. Nature 372:704-708(1994).
 [5] Valencia A., Chardin P., Wittinghofer A., Sander C. Biochemistry 30:4637-4648(1991).

(arf_2) ATP/GTP-binding site motif A (P-loop)

From sequence comparisons and crystallographic data analysis it has been shown

- 10 [1,2,3,4,5,6] that an appreciable proportion of proteins that bind ATP or GTP share a number
 of more or less conserved sequence motifs. The best conserved of these motifs is a glycine-
 rich region, which typically forms a flexible loop between a beta-strand and an alpha-helix.
 This loop interacts with one of the phosphate groups of the nucleotide. This sequence motif is
 generally referred to as the 'A' consensus sequence [1] or the 'P-loop' [5]. There are numerous
 15 ATP- or GTP-binding proteins in which the P-loop is found. A number of protein families for
 which the relevance of the presence of such motif has been noted are listed below: - ATP
 synthase alpha and beta subunits (see <PDOC00137>). - Myosin heavy chains. - Kinesin
 heavy chains and kinesin-like proteins (see <PDOC00343>). - Dynamins and dynamin-like
 proteins (see <PDOC00362>). - Guanylate kinase (see <PDOC00670>). - Thymidine kinase
 20 (see <PDOC00524>). - Thymidylate kinase (see <PDOC01034>). - Shikimate kinase (see
 <PDOC00868>). - Nitrogenase iron protein family (nifH/frxC) (see <PDOC00580>). - ATP-
 binding proteins involved in 'active transport' (ABC transporters) [7] (see <PDOC00185>). -
 DNA and RNA helicases [8,9,10]. - GTP-binding elongation factors (EF-Tu, EF-1alpha, EF-
 G, EF-2, etc.). - Ras family of GTP-binding proteins (Ras, Rho, Rab, Ral, Ypt1, SEC4, etc.).
 25 - Nuclear protein ran (see <PDOC00859>). - ADP-ribosylation factors family (see
 <PDOC00781>). - Bacterial dnaA protein (see <PDOC00771>). - Bacterial recA protein (see
 <PDOC00131>). - Bacterial recF protein (see <PDOC00539>). - Guanine nucleotide-binding
 proteins alpha subunits (Gi, Gs, Gt, G0, etc.). - DNA mismatch repair proteins mutS family
 (See <PDOC00388>). - Bacterial type II secretion system protein E (see <PDOC00567>). Not
 30 all ATP- or GTP-binding proteins are picked-up by this motif. A number of proteins escape
 detection because the structure of their ATP-binding site is completely different from that of
 the P-loop. Examples of such proteins are the E1-E2 ATPases or the glycolytic kinases. In
 other ATP- or GTP-binding proteins the flexible loop exists in a slightly different form; this

is the case for tubulins or protein kinases. A special mention must be reserved for adenylate kinase, in which there is a single deviation from the P-loop pattern: in the last position Gly is found instead of Ser or Thr.

5 Consensus pattern: [AG]-x(4)-G-K-[ST]-

[1] Walker J.E., Saraste M., Runswick M.J., Gay N.J. EMBO J. 1:945-951(1982).

[2] Moller W., Amons R. FEBS Lett. 186:1-7(1985).

[3] Fry D.C., Kuby S.A., Mildvan A.S. Proc. Natl. Acad. Sci. U.S.A. 83:907-911(1986).

10 [4] Dever T.E., Glynias M.J., Merrick W.C. Proc. Natl. Acad. Sci. U.S.A. 84:1814-1818(1987).

[5] Saraste M., Sibbald P.R., Wittinghofer A. Trends Biochem. Sci. 15:430-434(1990).

[6] Koonin E.V. J. Mol. Biol. 229:1165-1174(1993).

15 [7] Higgins C.F., Hyde S.C., Mimmack M.M., Gileadi U., Gill D.R., Gallagher M.P. J. Bioenerg. Biomembr. 22:571-592(1990).

[8] Hodgman T.C. Nature 333:22-23(1988) and Nature 333:578-578(1988) (Errata).

[9] Linder P., Lasko P., Ashburner M., Leroy P., Nielsen P.J., Nishi K., Schnier J., Slonimski P.P. Nature 337:121-122(1989).

20 [10] Gorbalenya A.E., Koonin E.V., Donchenko A.P., Blinov V.M. Nucleic Acids Res. 17:4713-4730(1989).

58. Arginase family signatures

25 The following enzymes have been shown [1] to be evolutionary related: - Arginase (EC 3.5.3.1), a ubiquitous enzyme which catalyzes the degradation of arginine to ornithine and urea [2]. - Agmatinase (EC 3.5.3.11) (agmatine ureohydrolase), a prokaryotic enzyme (gene speB) that catalyzes the hydrolysis of agmatine into putrescine and urea. -

Formiminoglutamase (EC 3.5.3.8) (formiminoglutamate hydrolase), a prokaryotic enzyme (gene hutG) that hydrolyzes N-formimino-glutamate into glutamate and formamide. -

30 Hypothetical proteins from methanogenic archaeobacteria. These enzymes are proteins of about 300 amino-acid residues. Three conserved regions that contain charged residues which are involved in the binding of the two manganese ions [3] can be used as signature patterns.-

Consensus pattern: [LIVMF SEQ ID NO:2)]-G-G-x-H-x-[LIVMT SEQ ID NO:1)]-[STAV
SEQ ID NO:105)]-x-[PAG]-x(3)-[GSTA SEQ ID NO:19)] [H binds manganese]-

Consensus pattern: [LIVM SEQ ID NO:4)](2)-x-[LIVMFY SEQ ID NO:18)]-D-[AS]-H-x-D
[The two D's and the H bind manganese]-

5 Consensus pattern: [ST]-[LIVMFY SEQ ID NO:18)]-D-[LIVM SEQ ID NO:4)]-D-x(3)-
[PAQ]-x(3)-P-[GSA]-x(7)-G [The two D's bind manganese]

[1] Ouzounis C., Kyripides N.C. J. Mol. Evol. 39:101-104(1994).

[2] Jenkinson C.P., Grody W.W., Cederbaum S.D. Comp. Biochem. Physiol. 114B:107-
10 132(196).

[3] Kanyo Z.F., Scolnick L.R., Ash D.E., Christianson D.W. Nature 383:554-557(1996).

59. (asp) Eukaryotic and viral aspartyl proteases active site

15 Aspartyl proteases, also known as acid proteases, (EC 3.4.23.-) are a widely distributed
family of proteolytic enzymes [1,2,3] known to exist invertebrates, fungi, plants, retroviruses
and some plant viruses. Aspartate proteases of eukaryotes are monomeric enzymes which
consist of two domains. Each domain contains an active site centered on a catalytic aspartyl
residue. The two domains most probably evolved from the duplication of an ancestral gene
20 encoding a primordial domain. Currently known eukaryotic aspartyl proteases are: -
Vertebrate gastric pepsins A and C (also known as gastricsin). - Vertebrate chymosin
(rennin), involved in digestion and used for making cheese. - Vertebrate lysosomal cathepsins
D (EC 3.4.23.5) and E (EC 3.4.23.34). - Mammalian renin (EC 3.4.23.15) whose function is
to generate angiotensin I from angiotensinogen in the plasma. - Fungal proteases such as
25 aspergillopepsin A (EC 3.4.23.18), candidapepsin (EC 3.4.23.24), mucoropepsin (EC
3.4.23.23) (mucor rennin), endothiapepsin (EC 3.4.23.22), polyporopepsin (EC 3.4.23.29),
and rhizopuspepsin (EC 3.4.23.21). - Yeast saccharopepsin (EC 3.4.23.25) (proteinase A)
(gene PEP4). PEP4 is implicated in posttranslational regulation of vacuolar hydrolases. -
Yeast barrier pepsin (EC 3.4.23.35) (gene BAR1); a protease that cleaves alpha-factor and
30 thus acts as an antagonist of the mating pheromone. - Fission yeast sxal which is involved in
degrading or processing the mating pheromones. Most retroviruses and some plant viruses,
such as badnaviruses, encode for an aspartyl protease which is an homodimer of a chain of
about 95 to 125 amino acids. In most retroviruses, the protease is encoded as a segment of

apolyprotein which is cleaved during the maturation process of the virus. It is generally part of the pol polyprotein and, more rarely, of the gagpolyprotein. Conservation of the sequence around the two aspartates of eukaryotic aspartyl proteases and around the single active site of the viral proteases allows us to develop a single signature pattern for both groups of protease.

5

Consensus pattern: [LIVMFGAC SEQ ID NO:106)]-[LIVMTADN SEQ ID NO:107)]-[LIVFSA SEQ ID NO:108)]-D-[ST]-G-[STAV SEQ ID NO:105)]-[STAPDENQ SEQ ID NO:109)]-x-[LIVMFSTNC SEQ ID NO:110)]-x-[LIVMFGTA SEQ ID NO:111)] [D is the active site residue]

10

Note: these proteins belong to families A1 and A2 in the classification of peptidases [4,E1

[1] Foltmann B. Essays Biochem. 17:52-84(1981).

[2] Davies D.R. Annu. Rev. Biophys. Chem. 19:189-215(1990).

[3] Rao J.K.M., Erickson J.W., Wlodawer A. Biochemistry 30:4663-4671(1991).

15

[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:105-120(1995).

60. (BIRA) Biotin repressor

[1] Wilson KP, Shewchuk LM, Brennan RG, Otsuka AJ, Matthews BW; Proc Natl Acad Sci USA 1992;89:9257-9261.

20

61. BTB/POZ domain

The BTB (for BR-C, ttk and bab) [1] or POZ (for Pox virus and Zinc finger)[2] domain is present near the N-terminus of a fraction of zinc finger

25

(zf-C2H2) proteins and in proteins that contain the Kelch motif

such as Kelch and a family of pox virus proteins. The BTB/POZ domain mediates

homomeric dimerisation and in some instances heteromeric dimerisation [2]. The structure of the dimerised PLZF BTB/POZ domain has been solved and consists of a tightly intertwined

30

homodimer. The central scaffolding of the protein is made up of a cluster of alpha-helices flanked by short beta-sheets at both the top and bottom of the molecule [3]. POZ domains from several zinc finger proteins have been shown to mediate transcriptional repression and

to interact with components of histone deacetylase co-repressor complexes including N-CoR and SMRT [4,5,6]. The POZ or BTB domain is also known as BR-C/Ttk or ZiN

[1] Zollman S, Godt D, Prive GG, Couderc JL, Laski FA; Proc Natl Acad Sci U S A 1994;91:10717-10721.

[2] Bardwell VJ, Treisman R; Genes Dev 1994;8:1664-1677.

[3] Ahmad KF, Engel CK, Prive GG; Proc Natl Acad Sci U S A 1998;95:12123-12128.

[4] Deweindt C, Albagli O, Bernardin F, Dhordain P, Quief S, Lantoine D, Kerckaert JP, Leprince D; Cell Growth Differ 1995;6:1495-1503.

[5] Huynh KD, Bardwell VJ; Oncogene 1998;17:2473-2484.

[6] Wong CW, Privalsky ML; J Biol Chem 1998;273:27695-27702.

62. (Bac GSPproteins) Bacterial type II secretion system protein D signature

A number of bacterial proteins, some of which are involved in a general secretion pathway (GSP) for the export of proteins (also called the type II pathway) [1 to 5], have been found to be evolutionary related. These proteins are listed below: - The 'D' protein from the GSP operon of: *Aeromonas* (gene *exeD*); *Erwinia* (gene *outD*); *Escherichia coli* (gene *yheF*), *Klebsiella pneumoniae* (gene *pulD*); *Pseudomonas aeruginosa* (gene *xcpQ*); *Vibrio cholerae* (gene *epsD*) and *Xanthomonas campestris* (gene *xpsD*). - *comE* from *Haemophilus influenzae*, involved in competence (DNA uptake). - *pilQ* from *Pseudomonas aeruginosa*, which is essential for the formation of the pili. - *hopQ* (*hopQ*) from *Escherichia coli*. - *hrpH* from *Pseudomonas syringae*, which is involved in the secretion of a proteinaceous elicitor of the hypersensitivity response in plants. - *hrpA1* from *Xanthomonas campestris* pv.

vesicatoria, which is also involved in the hypersensitivity response. - *mxuD* from *Shigella flexneri* which is involved in the secretion of the Ipa invasins which are necessary for penetration of intestinal epithelial cells. - *omc* from *Neisseria gonorrhoeae*. - *yssC* from *Yersinia enterocolitica* virulence plasmid pYV, which seems to be required for the export of the Yop virulence proteins. - The gpIV protein from filamentous phages such as *f1*, *ike*, or *m13*. GpIV is said to be involved in phage assembly and morphogenesis. These proteins all seem to start with a signal sequence and are thought to be integral proteins in the outer membrane. As a signature pattern a conserved region in the C-terminal section of these proteins has been selected

Consensus pattern: [GR]-[DEQKG SEQ ID NO:112)]-[STVM SEQ ID NO:113)]-[LIVMA
SEQ ID NO:30)](3)-[GA]-G-[LIVMFY SEQ ID NO:18)]-x(11)- [LIVM SEQ ID NO:4)]-P-
[LIVMFYWGS SEQ ID NO:114)]-[LIVMF SEQ ID NO:2)]-[GSAE SEQ ID NO:115)]-x-
5 [LIVM SEQ ID NO:4)]-P- [LIVMFYW SEQ ID NO:26)](2)-x(2)-[LV]-F

[1] Salmond G.P.C., Reeves P.J. Trends Biochem. Sci. 18:7-12(1993).

[2] Reeves P.J., Whitcombe D., Wharam S., Gibson M., Allison G., Bunce N., Barallon R.,
Douglas P., Mulholland V., Stevens S., Walker S., Salmond G.P.C. Mol. Microbiol. 8:443-
10 456(1993).

[3] Martin P.R., Hobbs M., Free P.D., Jeske Y., Mattick J.S. Mol. Microbiol. 9:857-
868(1993).

[4] Hobbs M., Mattick J.S. Mol. Microbiol. 10:233-243(1993).

[5] Genin S., Boucher C.A. Mol. Gen. Genet. 243:112-118(1994).

63. (Bac globin) Protozoan/cyanobacterial globins signature

Globins are heme-containing proteins involved in binding and/or transporting oxygen [1].

Almost all globins belong to a large family (see <PDOC00793>), the only exceptions are the
20 following proteins which form a family of their own[2,3]: - Monomeric hemoglobins from
the protozoan *Paramecium caudatum*, *Tetrahymena pyriformis* and *Tetrahymena*
thermophila. - Cyanoglobin from the cyanobacteria *Nostoc commune*. - Globins LI637 and
LI410 from the chloroplast of the alga *Chlamydomonas eugametos*. - *Mycobacterium*
tuberculosis hypothetical protein MtCY48.23. These proteins contain a conserved histidine
25 which could be involved in heme-binding. As a signature pattern, a conserved region that
ends with this residue was used

Consensus pattern: F-[LF]-x(5)-G-[PA]-x(4)-G-[KRA]-x-[LIVM SEQ ID NO:4)]-x(3)-H-

[1] Concise Encyclopedia Biochemistry, Second Edition, Walter de Gruyter, Berlin New-
York (1988).

[2] Takagi T. Curr. Opin. Struct. Biol. 3:413-418(1993).

[3] Couture M., Chamberland H., St-Pierre B., Lafontaine J., Guertin M.; Mol. Gen. Genet. 243:185-197(1994).

64. Band 7 protein family signature

Mammalian band 7 protein [1] (also known as 7.2B or stomatin) is an integral membrane phosphoprotein of red blood cells thought to regulate cation conductance by interacting with other proteins of the junctional complex of the membrane skeleton. Structurally, band 7 is evolutionary related to the following proteins: - *Caenorhabditis elegans* protein mec-2 [2]. Mec-2 positively regulates the activity of the putative mechanosensory transduction channel. It may links the mechanosensory channel and the microtubule cytoskeleton of the touch receptor neurons. - *Caenorhabditis elegans* proteins sto-1 to sto-4. - *Caenorhabditis elegans* protein unc-1. - *Escherichia coli* hypothetical protein ybbK. - *Mycobacterium tuberculosis* hypothetical protein MtCY277.09. - *Synechocystis* strain PCC 6803 hypothetical protein slr1128. - *Methanococcus jannaschii* hypothetical protein MJ0827. Structurally all these proteins consist of a short N-terminal domain which is followed by a transmembrane region and a variable size (from 170 to 350 residues) C-terminal domain. As a signature pattern, a conserved region located about 110 residues after the transmembrane domain was selected

Consensus pattern: R-x(2)-[LIV]-[SAN]-x(6)-[LIV]-D-x(2)-T-x(2)-W-G-[LIV]-[KRH]-[LIV]-x-[KR]-[LIV]-E-[LIV]-[KR]-

[1] Gallagher P.G., Forget B.G. *J. Biol. Chem.* 270:26358-26363(1995).

[2] Huang M., Gu G., Ferguson E.L., Chalfie M. *Nature* 378:292-295(1995).

65. Barwin domain signatures

Barwin [1] is a barley seed protein of 125 residues that binds weakly a chitin analog. It contains six cysteines involved in disulfide bonds, as shown in the following schematic representation.

+-----+ | ***** | *****

xxxxxxxxxxxxxxxxCxxxxxxxxCxxxxCxxxxxxxxCxxxxxxxxxxxxxxxxCxxxx || | +-----

-----+ +-----+'C': conserved cysteine involved in a disulfide bond.'*':

position of the patterns. Barwin is closely related to the following proteins: - Hevein, a wound-induced protein found in the latex of rubber trees. - HEL, an *Arabidopsis thaliana* hevein-like protein [2]. - Win1 and win2, two wound-induced proteins from potato. - Pathogenesis-related protein 4 from tobacco. Hevein and the win1/2 proteins consist of an N-terminal chitin-binding domain followed by a barwin-like C-terminal domain. Barwin and its related proteins could be involved in a defense mechanism in plants. As signature patterns, two highly conserved regions that contain some of the cysteines were selected

Consensus pattern: C-G-[KR]-C-L-x-V-x-N [The two C's are involved in disulfide bonds]-

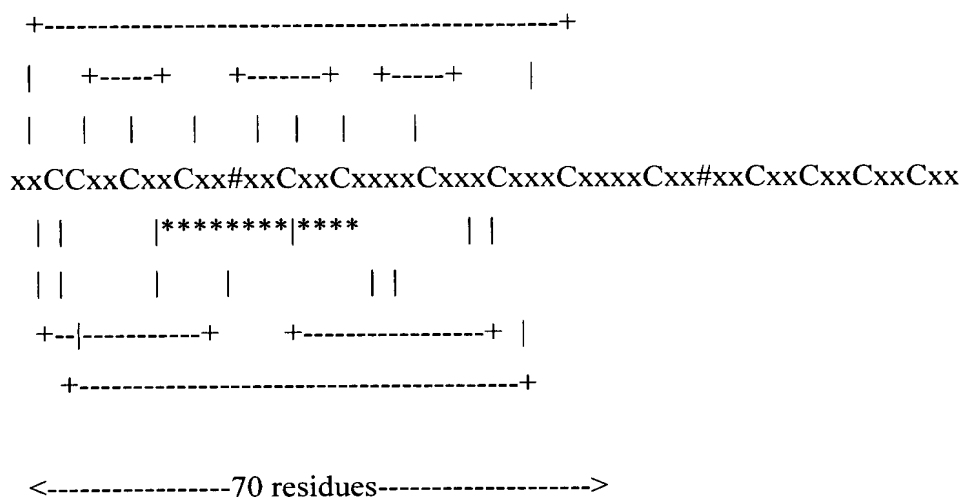
Consensus pattern: V-[DN]-Y-[EQ]-F-V-[DN]-C [C is involved in a disulfide bond]-

[1] Svensson B., Svendsen I., Hoejrup P., Roepstorff P., Ludvigsen S., Poulsen F.M. *Biochemistry* 31:8767-8770(1992).

[2] Potter S., Uknes S., Lawton K., Winter A.M., Chandler D., Dimaio J., Novitzky R., Ward E., Ryals J. *Mol. Plant Microbe Interact.* 6:680-685(1993).

66. (Bowman-Birk leg) Bowman-Birk serine protease inhibitors family signature

PROSITE cross-reference(s). The Bowman-Birk inhibitor family [1] is one of the numerous families of serine proteinase inhibitors. As it can be seen in the schematic representation, they have a duplicated structure and generally possess two distinct inhibitory sites:



'C': conserved cysteine involved in a disulfide bond.

'#': active site residue.

'*': position of the pattern.

5 These inhibitors are found in the seeds of all leguminous plants as well as in cereal grains. In cereals they exist in two forms, one of which is a duplication of the basic structure shown above [2]. The pattern that was developed to pick up sequences belonging to this family of inhibitors is in the central part of the domain and includes four cysteines.

10

Consensus pattern C-x(5,6)-[DENQKRHSTA SEQ ID NO:116)]-C-[PASTDH SEQ ID NO:117)]-[PASTDK SEQ ID NO:118)]-[ASTDV SEQ ID NO:119)]-C-[NDKS SEQ ID NO:120)]-[DEKRHSTA SEQ ID NO:121)]-C [The four C's are involved in disulfide bonds] Note this pattern can be found twice in some duplicated cereal inhibitors.

15

[1] Laskowski M., Kato I. Annu. Rev. Biochem. 49:593-626(1980).

[2] Tashiro M., Hashino K., Shiozaki M., Ibuki F., Maki Z. J. Biochem. 102:297-306(1987).

20 67. Pathogenesis-related protein Bet v I family signature

A number of plant proteins, which all seem to be involved in pathogen defense response, are structurally related [1,2,3]. These proteins are:

- Bet v I, the major pollen allergen from white birch. Bet v I is the main cause of type I allergic reactions in Europe, North America and USSR.
- 25 - Aln g I, the major pollen allergen from alder.
- Api G I, the major allergen from celery.
- Car b I, the major pollen allergen from hornbeam.
- Cor a I, the major pollen allergen from hazel.
- Mal d I, the major pollen allergen from apple.
- 30 - Asparagus wound-induced protein AoPR1.
- Kidney bean pathogenesis-related proteins 1 and 2.
- Parsley pathogenesis-related proteins PR1-1 and PR1-3.
- Pea disease resistance response proteins pI49, pI176 and DRRG49-C.

- Pea abscisic acid-responsive proteins ABR17 and ABR18.
- Potato pathogenesis-related proteins STH-2 and STH-21.
- Soybean stress-induced protein SAM22.

These proteins are thought to be intracellularly located. They contain from 155 to 160 amino acid residues. As a signature pattern, a conserved region located in the third quarter of these proteins has been selected

Consensus pattern: G-x(2)-[LIVMF SEQ ID NO:2]-x(4)-E-x(2)-[CSTAEN SEQ ID NO:122)]-x(8,9)-[GND]-G-[GS]- [CS]-x(2)-K-x(4)-[FY]-

[1] Breiteneder H., Pettenburger K., Bito A., Valenta R., Kraft D., Rumpold H., Scheiner O., Breitenbach M. EMBO J. 8:1935-1938(1989).

[2] Crowell D., John M.E., Russell D., Amasino R.M. Plant Mol. Biol. 18:459-466(1992).

[3] Warner S.A.J., Scott R., Draper J. Plant Mol. Biol. 19:555-561(1992).

68. bZIP transcription factors basic domain signature

The bZIP superfamily [1,2,] of eukaryotic DNA-binding transcription factors groups together proteins that contain a basic region mediating sequence-specific DNA-binding followed by a leucine zipper required for dimerization. This family is quite large, therefore only a partial list

of some representative members appears here. - Transcription factor AP-1, which binds selectively to enhancer elements in the cis control regions of SV40 and metallothionein IIA. AP-1, also known as c-jun, is the cellular homolog of the avian sarcoma virus 17 (ASV17) oncogene v-jun. - Jun-B and jun-D, probable transcription factors which are highly similar to jun/AP-1. - The fos protein, a proto-oncogene that forms a non-covalent dimer with c-jun. -

The fos-related proteins fra-1, and fos B. - Mammalian cAMP response element (CRE) binding proteins CREB, CREM, ATF-1, ATF-3, ATF-4, ATF-5, ATF-6 and LRF-1. - Maize Opaque 2, a trans-acting transcriptional activator involved in the regulation of the production of zein proteins during endosperm. - Arabidopsis G-box binding factors GBF1 to GBF4, Parsley CPRF-1 to CPRF-3, Tobacco TAF-1 and wheat EMBP-1. All these proteins bind the G-box promoter elements of many plant genes. - Drosophila protein Giant, which represses the expression of both the kruppel and knirps segmentation gap genes. - Drosophila Box B binding factor 2 (BBF-2), a transcriptional activator that binds to fat body-specific enhancers of alcohol dehydrogenase and yolk protein genes. - Drosophila segmentation protein

cap'n'collar (gene *cnc*), which is involved in head morphogenesis. - *Caenorhabditis elegans* *skn-1*, a developmental protein involved in the fate of ventral blastomeres in the early embryo. - Yeast *GCN4* transcription factor, a component of the general control system that regulates the expression of amino acid-synthesizing enzymes in response to amino acid starvation, and the related *Neurospora crassa* *cpc-1* protein. - *Neurospora crassa* *cys-3* which turns on the expression of structural genes which encode sulfur-catabolic enzymes. - Yeast *MET28*, a transcriptional activator of sulfur amino acids metabolism. - Yeast *PDR4* (or *YAP1*), a transcriptional activator of the genes for some oxygen detoxification enzymes. - Epstein-Barr virus trans-activator protein *BZLF1*.

Consensus pattern: [KR]-x(1,3)-[RKSAQ SEQ ID NO:123)]-N-x(2)-[SAQ](2)-x-[RKTAENQ SEQ ID NO:124)]-x-R-x-[RK]-

[1] Hurst H.C. *Protein Prog.* 2:105-168(1995).[2] Ellenberger T. *Curr. Opin. Struct. Biol.* 4:12-21(1994).

69. Biotin-requiring enzymes attachment site

Biotin, which plays a catalytic role in some carboxyl transfer reactions, is covalently attached, via an amide bond, to a lysine residue in enzymes requiring this coenzyme [1,2,3,4]. Such enzymes are:

- Pyruvate carboxylase (EC 6.4.1.1).
- Acetyl-CoA carboxylase (EC 6.4.1.2).
- Propionyl-CoA carboxylase (EC 6.4.1.3).
- Methylcrotonoyl-CoA carboxylase (EC 6.4.1.4).
- Geranoyl-CoA carboxylase (EC 6.4.1.5).
- Urea carboxylase (EC 6.3.4.6).
- Oxaloacetate decarboxylase (EC 4.1.1.3).
- Methylmalonyl-CoA decarboxylase (EC 4.1.1.41).
- Glutaconyl-CoA decarboxylase (EC 4.1.1.70).
- Methylmalonyl-CoA carboxyl-transferase (EC 2.1.3.1) (transcarboxylase).

Sequence data reveal that the region around the biocytin (biotin-lysine) residue is well conserved and can be used as a signature pattern.

Consensus pattern[GN]-[DEQTR SEQ ID NO:125)]-x-[LIVMFY SEQ ID NO:18)]-x(2)-
 [LIVM SEQ ID NO:4)]-x-[AIV]-M-K-[LMAT SEQ ID NO:126)]-x(3)-[LIVM SEQ ID
 NO:4)]-x-[SAV] [K is the biotin attachment site] Note the domain around the biotin-binding
 5 lysine residue is evolutionary related to that around the lipoyl-binding lysine residue of 2-oxo
 acid dehydrogenase acyltransferases

[1] Knowles J.R. Annu. Rev. Biochem. 58:195-221(1989).

10 [2] Samols D., Thronton C.G., Murtif V.L., Kumar G.K., Haase F.C., Wood H.G. J. Biol.
 Chem. 263:6461-6464(1988).

[3] Goss N.H., Wood H.G. Meth. Enzymol. 107:261-278(1984).

[4] Shenoy B.C., Xie Y., Park V.L., Kumar G.K., Beegen H., Wood H.G., Samols D. J. Biol.
 Chem. 267:18407-18412(1992).

15 2-oxo acid dehydrogenases acyltransferase component lipoyl binding site

The 2-oxo acid dehydrogenase multienzyme complexes [1,2] from bacterial and
 eukaryotic sources catalyze the oxidative decarboxylation of 2-oxo acids to
 the corresponding acyl-CoA. The three members of this family of multienzyme
 complexes are:

- 20 - Pyruvate dehydrogenase complex (PDC).
 - 2-oxoglutarate dehydrogenase complex (OGDC).
 - Branched-chain 2-oxo acid dehydrogenase complex (BCOADC).

These three complexes share a common architecture: they are composed of
 multiple copies of three component enzymes - E1, E2 and E3. E1 is a thiamine
 25 pyrophosphate-dependent 2-oxo acid dehydrogenase, E2 a dihydrolipamide
 acyltransferase, and E3 an FAD-containing dihydrolipamide dehydrogenase.
 E2 acyltransferases have an essential cofactor, lipoic acid, which is
 covalently bound via an amide linkage to a lysine group. The E2 components of
 OGCD and BCOACD bind a single lipoyl group, while those of PDC bind either one
 30 (in yeast and in *Bacillus*), two (in mammals), or three (in *Azotobacter* and in
Escherichia coli) lipoyl groups [3].

In addition to the E2 components of the three enzymatic complexes described
 above, a lipoic acid cofactor is also found in the following proteins:

- H-protein of the glycine cleavage system (GCS) [4]. GCS is a multienzyme complex of four protein components, which catalyzes the degradation of glycine. H protein shuttles the methylamine group of glycine from the P protein to the T protein. H-protein from either prokaryotes or eukaryotes binds a single lipoic group.
- Mammalian and yeast pyruvate dehydrogenase complexes differ from that of other sources, in that they contain, in small amounts, a protein of unknown function - designated protein X or component X. Its sequence is closely related to that of E2 subunits and seems to bind a lipoic group [5].
- Fast migrating protein (FMP) (gene acoC) from *Alcaligenes eutrophus* [6]. This protein is most probably a dihydrolipamide acyltransferase involved in acetoin metabolism.

A signature pattern was developed which allows the detection of the lipoyl-binding site.

Consensus pattern[GN]-x(2)-[LIVF SEQ ID NO:127)]-x(5)-[LIVFC SEQ ID NO:128)]-x(2)-[LIVFA SEQ ID NO:129)]-x(3)-K-[STAIV SEQ ID NO:130)]-[STAVQDN SEQ ID NO:131)]-x(2)-[LIVMFS SEQ ID NO:132)]-x(5)-[GCN]-x-[LIVMFY SEQ ID NO:18)] [K is the lipoyl-binding site] Note the domain around the lipoyl-binding lysine residue is evolutionary related to that around the biotin-binding lysine residue of biotin requiring enzymes

[1] Yeaman S.J. Biochem. J. 257:625-632(1989).

[2] Yeaman S.J. Trends Biochem. Sci. 11:293-296(1986).

[3] Russel G.C., Guest J.R. Biochim. Biophys. Acta 1076:225-232(1991).

[4] Fujiwara K., Okamura-Ikeda K., Motokawa Y. J. Biol. Chem. 261:8836-8841(1986).

[5] Behal R.H., Browning K.S., Hall T.B., Reed L.J. Proc. Natl. Acad. Sci. U.S.A. 86:8732-8736(1989).

[6] Priefert H., Hein S., Krueger N., Zeh K., Schmidt B., Steinbuechel A. J. Bacteriol.

173:4056-4071(1991).

Some isozymes of protein kinase C (PKC) [1,2] contain a domain, known as C2, of about 116 amino-acid residues which is located between the two copies of the C1 domain (that bind phorbol esters and diacylglycerol) (see <PDOC00379>) and the protein kinase catalytic domain (see <PDOC00100>). Regions with significant homology [3,E1] to the C2-domain have been found in the following proteins:

- PKC isoforms alpha, beta and gamma and Drosophila isoforms PKC1 and PKC2.
- PKC isoforms delta, epsilon and eta, Caenorhabditis elegans kin-13 and yeast PKC1 have a C2-like domain at the N-terminal extremity [4].
- Yeast cAMP dependent protein kinase SCH9 contains a C2-like domain.
- Mammalian phosphatidylinositol-specific phospholipase C (PI-PLC) (see <PDOC50007>) isoforms beta, gamma and delta as well as several non-mammalian PI-PLCs have a C2-like domain C-terminal of the catalytic domain.
- Mammalian and plants phosphatidylinositol-3-kinase have a C2-like domain in the central region of the 110 Kd catalytic subunit.
- Yeast phosphatidylserine-decarboxylase 2 (gene PSD2) contains a C2 domain in its central region.
- Cytosolic phospholipase D from plants and cytosolic phospholipase A2 have a C2-like domain at their N-terminus.
- Synaptotagmins (p65). This is a family of related synaptic vesicle proteins that bind acidic phospholipids and that may have a regulatory role in the membrane interactions during trafficking of synaptic vesicles at the active zone of the synapse. All isoforms of synaptotagmins have two copies of the C2 domain in their C-terminal region.
- Rabphilin-3A, a synaptic protein contains two C2 domains.
- Caenorhabditis elegans protein unc-13 whose function is not known. Unc-13 has a C2 domain in its central part and a C2-like domain at the C-terminus.
- rasGAP and the breakpoint cluster protein bcr have a C2-domain C-terminal of a PH-domain.
- Yeast protein BUD2 (or CLA2) has a C2-domain in the central region.
- Yeast protein RSP5 and human protein NEDD-4, both proteins also contain WW domains (see <PDOC50020>).
- Perforin (see <PDOC00251>) has a C2 domain at the C-terminus. It is the only extracellular protein known to contain a C2 domain.
- Yeast hypothetical protein YML072C has a C2 domain.

- Yeast hypothetical protein YNL087W has three C2 domains.
- Caenorhabditis elegans hypothetical protein F37A4.7 has two C2 domains.

The C2 domain is thought to be involved in calcium-dependent phospholipid binding [5].

Since domains related to the C2 domain are also found in proteins that do not bind calcium,

5 other putative functions for the C2 domain like e.g. binding to inositol-1,3,4,5-tetraphosphate have been suggested [6]. Recently, the 3D structure of the first C2 domain of

synaptotagmin has been reported [7], the domain forms an eight-stranded beta sandwich. The

signature pattern that has been developed for the C2 domain is located in a conserved part of

that domain, the connecting loop between beta strands 2 and 3. A profile has been

10 developed for the C2 domain that covers the total domain.

-Consensus pattern: [ACG]-x(2)-L-x(2,3)-D-x(1,2)-[NGSTLIF SEQ ID NO:133)]-[GTMR
SEQ ID NO:134)]-x-[STAP SEQ ID NO:135)]-D-[PA]-[FY]

-Note: this documentation entry is linked to both a signature pattern and a profile. As the
15 profile is much more sensitive than the pattern, you should use it if you have access to the
necessary software tools to do so.

[1]Medline: 96367095 Extending the C2 domain family: C2s in PKCs delta, epsilon, eta and
20 theta, phospholipases, GAPs and perforin. Ponting CP, Parker PJ; Protein Sci 1996;5:162-
166.

[1] Azzi A., Boscoboinik D., Hensey C. Eur. J. Biochem. 208:547-557(1992).

[2] Stabel S. Semin. Cancer Biol. 5:277-284(1994).

[3] Brose N., Hofmann K.O., Hata Y., Suedhof T.C. J. Biol. Chem. 270:25273-25280(1995).

[4] Sossin W.S., Schwartz J.H. Trends Biochem. Sci. 18:207-208(1993).

25 [5] Davletov B.A., Suedhof T.C. J. Biol. Chem. 268:26386-26390(1993).

[6] Fukuda M., Aruga J., Niinobe M., Aimoto S., Mikoshiba K. J. Biol. Chem. 269:29206-
29211(1994).

[6] Sutton R.B., Davletov B.A., Berghuis A.M., Suedhof T.C., Sprang S.R. Cell 80:929-
938(1995).

30

71. CAP (CAP protein) Number of members: 11

In budding and fission yeasts the CAP protein is a bifunctional protein whose N-terminal domain binds to adenylyl cyclase, thereby enabling that enzyme to be activated by upstream regulatory signals, such as Ras. The function of the C-terminal domain is less clear, but it is required for normal cellular morphology and growth control [1]. CAP is conserved in higher eukaryotic organisms where its function is not yet clear [2].

Structurally, CAP is a protein of 474 to 551 residues which consist of two domains separated by a proline-rich hinge. Two signature patterns, one corresponding to a conserved region in the N-terminal extremity and the other to a C-terminal region have been developed.

-Consensus pattern: [LIVM SEQ ID NO:4])(2)-x-R-L-[DE]-x(4)-R-L-E

-Consensus pattern: D-[LIVMFY SEQ ID NO:18)]-x-E-x-[PA]-x-P-E-Q-[LIVMFY SEQ ID NO:18)]-K

[1] Kawamukai M., Gerst J., Field J., Riggs M., Rodgers L., Wigler M., Young D. Mol. Biol. Cell 3:167-180(1992).

[2] Yu G., Swiston J., Young D. J. Cell Sci. 107:1671-1678(1994).

72. CAP_GLY (CAP-Gly domain)

CAP stands for cytoskeleton-associated proteins. Swiss:P39937 may be a member but has not been included. It has a weak match to the family between residues 22-67. Number of members: 24

[1] Medline: 93242656. Sequence homologies between four cytoskeleton-associated proteins.

Riehemann K, Sorg C; Trends Biochem Sci 1993;18:82-83.

It has been shown [1] that some cytoskeleton-associated proteins (CAP) share the presence of a conserved, glycine-rich domain of about 42 residues, called here CAP-Gly. Proteins known to contain this domain are listed below.

- Restin (also known as cytoplasmic linker protein-170 or CLIP-170), a 160 Kd protein associated with intermediate filaments and that links endocytic vesicles to microtubules. Restin contains two copies of the CAP-Gly domain.

- Vertebrate dynactin (150 Kd dynein-associated polypeptide; DAP) and *Drosophila* glued, a major component of activator I, a 20S polypeptide complex that stimulates dynein-mediated vesicle transport.

- Yeast protein BIK1 which seems to be required for the formation or stabilization of microtubules during mitosis and for spindle pole body fusion during conjugation.

- Yeast protein NIP100 (NIP80).

- Human protein CKAP1/TFCB, *Schizosaccharomyces pombe* protein alp11 and *Caenorhabditis elegans* hypothetical protein F53F4.3. These proteins contain a N-terminal ubiquitin domain (see <PDOC00271>) and a C-terminal CAP-Gly domain.

- *Caenorhabditis elegans* hypothetical protein M01A8.2.

- Yeast hypothetical protein YNL148c.

Structurally, these proteins are made of three distinct parts: an N-terminal section that is most probably globular and contains the CAP-Gly domain, a large central region predicted to be in an alpha-helical coiled-coil conformation and, finally, a short C-terminal globular domain. The signature for the CAP-Gly domain corresponds to the first 32 residues of the domain and includes five of the six conserved glycines.

-Consensus pattern: G-x(8,10)-[FYW]-x-G-[LIVM SEQ ID NO:4)]-x-[LIVMFY SEQ ID NO:18)]-x(4)-G-K-[NH]-x-G-[STAR SEQ ID NO:136)]-x(2)-G-x(2)-[LY]-F

[1] Riehemann K., Sorg C. Trends Biochem. Sci. 18:82-83(1993).

73. (CBD 1)

Cellulose-binding domain, fungal type

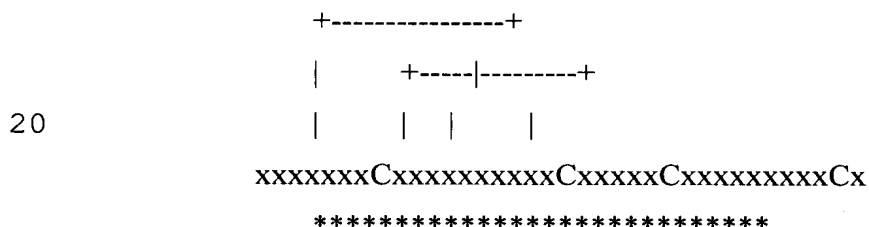
The microbial degradation of cellulose and xylans requires several types of enzymes such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases (EC 3.2.1.8) [1].

Structurally, cellulases and xylanases generally consist of a catalytic domain joined to a cellulose-binding domain (CBD) by a short linker sequence rich in proline and/or hydroxy-amino acids.

The CBD of a number of fungal cellulases has been shown to consist of 36 amino acid residues. Enzymes known to contain such a domain are:

- Endoglucanase I (gene egl1) from *Trichoderma reesei*.
- 5 - Endoglucanase II (gene egl2) from *Trichoderma reesei*.
- Endoglucanase V (gene egl5) from *Trichoderma reesei*.
- Exocellobiohydrolase I (gene CBHI) from *Humicola grisea*, *Neurospora crassa*,
Phanerochaete chrysosporium, *Trichoderma reesei*, and *Trichoderma viride*.
- Exocellobiohydrolase II (gene CBHII) from *Trichoderma reesei*.
- 10 - Exocellobiohydrolase 3 (gene cel3) from *Agaricus bisporus*
- Endoglucanases B, C2, F and K from *Fusarium oxysporum*.

The CBD domain is found either at the N-terminal (Cbh-II or egl2) or at the C-terminal extremity (Cbh-I, egl1 or egl5) of these enzymes. As it is shown in the following schematic representation, there are four conserved cysteines in this type of CBD domain, all involved in disulfide bonds.



'C': conserved cysteine involved in a disulfide bond.

25 '*' : position of the pattern.

Such a domain has also been found in a putative polysaccharide binding protein from the red alga, *Porphyra purpurea* [2]. Structurally, this protein consists of four tandem repeats of the CBD domain.

30 Consensus pattern C-G-G-x(4,7)-G-x(3)-C-x(5)-C-x(3,5)-[NHG]-x-[FYWM SEQ ID NO:137)]-x(2)-Q-C [The four C's are involved in disulfide bonds] Sequences known to belong to this class detected by the pattern ALL.

[1] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev. 55:303-315(1991).

[2] Liu Q., der Meer J.P., Reith M.E.

5

74. CBS domain. 3D Structure found as a subdomain in TIM barrel of inosine-. CBS domain web page. CBS domains are small intracellular modules mostly found in 2 or four copies within a protein. CBS domains are found in cystathionine-beta-synthase (CBS) where mutations lead to homocystinuria. Two CBS domains are found in inosine-monophosphate dehydrogenase from all species, however the CBS domains are not needed for activity. Two CBS domains are found in intracellular loops of several chloride channels. Mutations in this domain of Swiss:P35520 lead to homocystinuria.

Number of members: 414

15

[1]Medline: 97172695 The structure of a domain common to archaeobacteria and the homocystinuria disease protein. Bateman A; Trends Biochem Sci 1997;22:12-13.

[2]Medline: 96279836 Structure and mechanism of inosine monophosphate dehydrogenase in complex with the immunosuppressant mycophenolic-acid. Sintchak MD, Fleming MA, Futer O, Raybuck SA, Chambers SP, Caron PR, Murcko MA, Wilson KP; Cell 1996;85:921-930.

20

Discovery of CBS domain.

[3]Medline: 97259972 CBS domains in ClC chloride channels implicated in myotonia and nephrolithiasis (kidney stones). Ponting CP; J Mol Med 1997;75:160-163.

25

75. CDP-OH_P_transf (CDP-alcohol phosphatidyltransferase)

All of these members have the ability to catalyze the displacement of CMP from a CDP-alcohol by a second alcohol with formation of a phosphodiester bond and concomitant breaking of a phosphoride anhydride bond. Number of members: 32

30

A number of phosphatidyltransferases, which are all involved in phospholipid biosynthesis and that share the property of catalyzing the displacement of CMP from a CDP-alcohol by a

second alcohol with formation of a phosphodiester bond and concomitant breaking of a phosphoride anhydride bond share a conserved sequence region [1,2]. These enzymes are:

- Ethanolaminephosphotransferase (EC 2.7.8.1) from yeast (gene EPT1).
- Diacylglycerol cholinephosphotransferase (EC 2.7.8.2) from yeast (gene CPT1).
- 5 - Phosphatidylglycerophosphate synthase (EC 2.7.8.5) (CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase) from bacteria (gene pgsA).
- Phosphatidylserine synthase (EC 2.7.8.8) (CDP-diacylglycerol--serine O-phosphatidyltransferase) from yeast (gene CHO1) and from *Bacillus subtilis* (gene pssA).
- Phosphatidylinositol synthase (EC 2.7.8.11) (CDP-diacylglycerol--inositol 3-phosphatidyltransferase) from yeast (gene PIS).

These enzymes are proteins of from 200 to 400 amino acid residues. The conserved region contains three aspartic acid residues and is located in the N-terminal section of the sequences.

15 -Consensus pattern: D-G-x(2)-A-R-x(8)-G-x(3)-D-x(3)-D

[1]Medline: 97075020 Two-dimensional ¹H-NMR of transmembrane peptides from *Escherichia coli* phosphatidylglycerophosphate synthase in micelles. Morein S, Trouard TP, Hauksson JB, Rilfors L, Arvidson G, Lindblom G; *Eur J Biochem* 1996;241:489-497.

20 [1] Nikawa J.-I., Kodaki T., Yamashita S.

J. Biol. Chem. 262:4876-4881(1987).

[2] Hjelmstad R.H., Bell R.M.

J. Biol. Chem. 266:5094-5134(1991).

25 76. CHOD (Cholesterol oxidase) Members of the GMC oxidoreductase family. Number of members: 3

[1]Medline: 94032271. Crystal structure of cholesterol oxidase complexed with a steroid substrate: implications for flavin adenine dinucleotide dependent alcohol oxidases. Li J, Vrielink A, Brick P, Blow DM; *Biochemistry* 1993;32:11507-11515.

The following FAD flavoproteins oxidoreductases have been found [1,2] to be evolutionary related. These enzymes, which are called 'GMC oxidoreductases', are listed below.

- Glucose oxidase (EC 1.1.3.4) (GOX) from *Aspergillus niger*. Reaction catalyzed: glucose + oxygen -> delta-luconolactone + hydrogen peroxide.

5 - Methanol oxidase (EC 1.1.3.13) (MOX) from fungi. Reaction catalyzed: methanol + oxygen -> acetaldehyde + hydrogen peroxide.

- Choline dehydrogenase (EC 1.1.99.1) (CHD) from bacteria. Reaction catalyzed: choline + unknown acceptor -> betaine acetaldehyde + reduced acceptor.

10 - Glucose dehydrogenase (GLD) (EC 1.1.99.10) from *Drosophila*. Reaction catalyzed: glucose + unknown acceptor -> delta-gluconolactone + reduced acceptor.

- Cholesterol oxidase (CHOD) (EC 1.1.3.6) from *Brevibacterium sterolicum* and *Streptomyces* strain SA-COO. Reaction catalyzed: cholesterol + oxygen -> cholest-4-en-3-one + hydrogen peroxide.

15 - AlkJ [3], an alcohol dehydrogenase from *Pseudomonas oleovorans*, which converts aliphatic medium-chain-length alcohols into aldehydes. This family also includes a lyase:

- (R)-mandelonitrile lyase (EC 4.1.2.10) (hydroxynitrile lyase) from plants [4], an enzyme involved in cyanogenesis, the release of hydrogen cyanide from injured tissues.

20 These enzymes are proteins of size ranging from 556 (CHD) to 664 (MOX) amino acid residues which share a number of regions of sequence similarities. One of these regions, located in the N-terminal section, corresponds to the FAD ADP- binding domain. The function of the other conserved domains is not yet known; two of these domains have been selected as signature patterns. The first one is located in the N-terminal section of these enzymes, about 50 residues after the ADP-binding domain, while the second one is located in the central section.

25 -Consensus pattern: [GA]-[RKN]-x-[LIV]-G(2)-[GST](2)-x-[LIVM SEQ ID NO:4)]-N-x(3)-[FYWA SEQ ID NO:138)]- x(2)-[PAG]-x(5)-[DNESH SEQ ID NO:139)]

-Consensus pattern: [GS]-[PSTA SEQ ID NO:140)]-x(2)-[ST]-P-x-[LIVM SEQ ID NO:4)](2)-x(2)-S-G-[LIVM SEQ ID NO:4)]-G

30 [1] Cavener D.R. J. Mol. Biol. 223:811-814(1992).

[2] Henikoff S., Henikoff J.G. Genomics 19:97-107(1994).

[3] van Beilen J.B., Eggink G., Enequist H., Bos R., Witholt B. Mol. Microbiol. 6:3121-3136(1992).

[4] Cheng I.P., Poulton J.E. Plant Cell Physiol. 34:1139-1143(1993).

5

77. CKS (Cyclin-dependent kinase regulatory subunit) Number of members: 11. Cyclin-dependent kinases (CDK) are protein kinases which associate with cyclins to regulate eukaryotic cell cycle progression. The most well known CDK is p34-cdc2 (CDC28 in yeast) which is required for entry into S-phase and mitosis. CDK's bind to a regulatory subunit which is essential for their biological function. This regulatory subunit is a small protein of 79 to 150 residues. In yeast (gene CKS1) and in fission yeast (gene suc1) a single isoform is known, while mammals have two highly related isoforms. It has been shown [1] that these CDK regulatory subunits assemble as an hexamer which then acts as a hub for the oligomerization of six CDK catalytic subunits. The sequence of CDK regulatory subunits are highly conserved therefore, the two most conserved regions have been used as signature patterns.

10

15

-Consensus pattern: Y-S-x-[KR]-Y-x-[DE](2)-x-[FY]-E-Y-R-H-V-x-[LV]-[PT]-[KRP]

-Consensus pattern: H-x-P-E-x-H-[IV]-L-L-F-[KR]

20

[1] Parge H.E., Arvai A.S., Murtari D.J., Reed S.I., Tainer J.A. Science 262:387-395(1993).

78. CK_II_beta (Casein kinase II regulatory subunit)

25

30

Number of members: 16. Casein kinase II (CK-2) [1] is an ubiquitous eukaryotic serine/threonine protein kinase which is found both in the cytoplasm and the nucleus and whose substrates are numerous. It generally phosphorylates Ser or Thr at the N-terminal of stretch of acidic residues (see <PDOC00006>). CK-2 exists as an heterotetramer composed of two catalytic subunits (alpha) and two regulatory subunits (beta). In most species there are two closely related isoforms of the catalytic subunit: alpha and alpha'. Some species, such as fungi and plants, express two forms of regulatory subunits: beta and beta'. The exact function of the regulatory subunit is not yet known. It is a highly conserved protein of about 25 Kd that contains, in its central section, a cysteine-rich motif that could

be involved in binding a metal such as zinc [2]. This region has been used as a signature pattern.

-Consensus pattern: C-P-x-[LIVMY SEQ ID NO:141)]-x-C-x(5)-[LI]-P-[LIVMC SEQ ID NO:142)]-G-x(9)-V-[KR]-x(2)-C-P-x-C

[1] Allende J.E., Allende C.C. FASEB J. 9:313-323(1995).

[2] Reed J.C., Bidwai A.P., Glover C.V.C. J. Biol. Chem. 269:18192-18200(1994).

79. CLP_protease (Clp protease)

These proteins belong to family S14 in the classification of peptidases.

-!- The Clp protease has an active site catalytic triad. In E. coli Clp protease, ser-111, his-136 and asp-185 form the catalytic triad.

-!- Swiss:P48254 has lost all of these active site residues and is therefore inactive.

-!- Swiss:P42379 contains two large insertions, Swiss:P42380 contains one large insertion.

Number of members: 38

The endopeptidase Clp (EC 3.4.21.92) from Escherichia coli cleaves peptides in various proteins in a process that requires ATP hydrolysis [1,2]. Clp is a dimeric protein which consists of a proteolytic subunit (gene clpP) and either of two related ATP-binding regulatory subunits (genes clpA and clpX). ClpP is a serine protease which has a chymotrypsin-like activity. Its catalytic activity seems to be provided by a charge relay system similar to that of the trypsin family of serine proteases, but which evolved by independent convergent evolution. Proteases highly similar to ClpP have been found to be encoded in the genome of the chloroplast of plants and seem to be also present in other eukaryotes. The sequences around two of the residues involved in the catalytic triad (a serine and a histidine) are highly conserved and can be used as signature patterns specific to that category of proteases.

-Consensus pattern: T-x(2)-[LIVMF SEQ ID NO:2)]-G-x-A-[SAC]-S-[MSA]-[PAG]-[STA]
[S is the active site residue]

-Consensus pattern: R-x(3)-[EAP]-x(3)-[LIVMFYT SEQ ID NO:143)]-M-[LIVM SEQ ID NO:4)]-H-Q-P [H is the active site residue]

[1]Medline: 98050920. The structure of ClpP at 2.3 angstroms resolution suggests a model for ATP-dependent proteolysis. Wang J, Hartling JA, Flanagan JM; Cell 1997;91:447-456.

[1] Maurizi M.R., Clark W.P., Kim S.-H., Gottesman S. J. Biol. Chem. 265:12546-12552(1990).

[2] Gottesman S., Maurizi M.R. Microbiol. Rev. 56:592-621(1992).

[3] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).

80. CNG_membrane (Transmembrane region cyclic Nucleotide Gated Channel)

[1]Medline: 94224763. Cyclic nucleotide-gated channels: an expanding new family of ion channels. Yau KW; Proc Natl Acad Sci USA 1994;91:3481-3483.

This family is found to the N-terminus of the cNMP_binding. Number of members: 56.

Proteins that bind cyclic nucleotides (cAMP or cGMP) share a structural domain of about

120 residues [1-3]. The best studied of these proteins is the prokaryotic catabolite gene activator (also known as the cAMP receptor protein) (gene crp) where such a domain is known to be composed of three alpha-helices and a distinctive eight-stranded, antiparallel beta-barrel structure. Such a domain is known to exist in the following proteins:

- Prokaryotic catabolite gene activator protein (CAP).

- cAMP- and cGMP-dependent protein kinases (cAPK and cGPK). Both types of kinases contains two tandem copies of the cyclic nucleotide-binding domain. The cAPK's are composed of two different subunits: a catalytic chain and a regulatory chain which contains both copies of the domain. The cGPK's are single chain enzymes that include the two copies of the domain in their N-terminal section. The nucleotide specificity of cAPK and cGPK is due to an amino acid in the conserved region of beta-barrel 7: a threonine that is invariant in cGPK is an alanine in most cAPK.

- Vertebrate cyclic nucleotide-gated ion-channels. Two such cations channels have been fully characterized. One is found in rod cells where it plays a role in visual signal transduction. It specifically binds to cGMP leading to an opening of the channel and thereby causing a depolarization of rod photoreceptors. In olfactory epithelium a similar, cAMP-binding, channel plays a role in odorant signal transduction. There are six invariant amino acids in this domain, three of which are glycine residues that are thought to be essential for maintenance of the structural integrity of the beta-barrel. Two signature

patterns have been developed for this domain. The first pattern is located within beta-barrels and 3 and contains the first two conserved Gly. The second pattern is located within beta-barrels 6 and 7 and contains the third conserved Gly as well as the three other invariant residues.

5

-Consensus pattern: [LIVM SEQ ID NO:4)]-[VIC]-x(2)-G-[DENQTA SEQ ID NO:144)]-x-[GAC]-x(2)-[LIVMFY SEQ ID NO:18)](4)-x(2)-G

-Consensus pattern: [LIVMF SEQ ID NO:2)]-G-E-x-[GAS]-[LIVM SEQ ID NO:4)]-x(5,11)-R-[STAQ SEQ ID NO:145)]-A-x-[LIVMA SEQ ID NO:30)]-x-[STACV SEQ ID NO:146)]

10

[1] Weber I.T., Shabb J.B., Corbin J.D. Biochemistry 28:6122-6127(1989).

[2] Kaupp U.B. Trends Neurosci. 14:150-157(1991).

[3] Shabb J.B., Corbin J.D. J. Biol. Chem. 267:5723-5726(1992).

15

81. COX10_ctaB_cyoE (Cytochrome c oxidase assembly factor)

[1]Medline: 95191390

Biosynthesis and functional role of haem O and haem A

Mogi T, Saiki K, Anraku Y; Mol Microbiol 1994;14:391-398.

20

Cytochrome c oxidase is a multi subunit enzyme. The complexity of this enzyme requires assistance in building the complex.

This is carried out by the Cytochrome c oxidase assembly factor.

Number of members: 31

25

Cytochrome c oxidase is an oligomeric enzymatic complex which seems to require the aid of a number of proteins that either act as chaperonins to help the subunits of the enzyme to fold correctly, or assist in the assembly of the metal centers [1]. One of these subunits is known as COX10 in yeast and as ctaB [2] in aerobic prokaryotes. It is evolutionary related to cyoE protein from the Escherichia coli cytochrome O terminal oxidase complex.

30

These proteins probably contain [3] seven transmembrane segments. The most conserved region is located in a loop between the second and third of these

segments and has been selected as a signature pattern.

-Consensus pattern: [ED]-x-D-x(2)-M-x-R-T-x(2)-R-x(4)-G

- 5 [1] Nobrega M.P., Nobrega F.G., Tzagoloff A.
J. Biol. Chem. 265:14220-14226(1990).
[2] Cao J., Hosler J., Shapleigh J., Revzin A., Ferguson-Miller S.
J. Biol. Chem. 267:24273-24278(1992).
[3] Chepuri V., Gennis R.B.
10 J. Biol. Chem. 265:12978-12986(1990).

82. COX3 (Cytochrome c oxidase subunit III)

This family corresponds to chains c and p.

- 15 [1]Medline: 96216288
The whole structure of the 13-subunit oxidized cytochrome c
oxidase at 2.8 Å. Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H,
Shinzawa-Itoh K, Nakashima R, Yaono R, Yoshikawa S; Science 1996;272:1136-1144.
Number of members: 224

20

83. COX5B (Cytochrome c oxidase subunit Vb)

[1]

Medline: 96216288

- 25 The whole structure of the 13-subunit oxidized cytochrome c
oxidase at 2.8 Å.
Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H,
Shinzawa-Itoh K, Nakashima R, Yaono R, Yoshikawa S;
Science 1996;272:1136-1144.
30 This family consists of chains F and S
Number of members: 10

Cytochrome c oxidase (EC 1.9.3.1) [1] is an oligomeric enzymatic complex which

is a component of the respiratory chain complex and is involved in the transfer of electrons from cytochrome c to oxygen. In eukaryotes this enzyme complex is located in the mitochondrial inner membrane; in aerobic prokaryotes it is found in the plasma membrane. In addition to the three large subunits that form the catalytic center of the enzyme complex there are, in eukaryotes, a variable number of small polypeptidic subunits. One of these subunits which is known as Vb in mammals, V in slime mold and IV in yeast, binds a zinc atom. The sequence of subunit Vb is well conserved and includes three conserved cysteines that are thought to coordinate the zinc ion [2]. Two of these cysteines are clustered in the C-terminal section of the subunit; this region has been selected as a signature pattern.

-Consensus pattern: [LIVM SEQ ID NO:4](2)-[FYW]-x(10)-C-x(2)-C-G-x(2)-[FY]-K-L
[The two C's probably bind zinc]

[1] Capaldi R.A., Malatesta F., Darley-Usmar V.M.

Biochim. Biophys. Acta 726:135-148(1983).

[2] Rizzuto R., Sandona D., Brini M., Capaldi R.A., Bisson R.

Biochim. Biophys. Acta 1129:100-104(1991).

84. COesterase (Carboxylesterases)

Cholinesterase pages

The prints entry is specific to acetylcholinesterase

Number of members: 273

Higher eukaryotes have many distinct esterases. Among the different types are those which act on carboxylic esters (EC 3.1.1.-). Carboxyl-esterases have been classified into three categories (A, B and C) on the basis of differential patterns of inhibition by organophosphates. The sequence of a number of type-B carboxylesterases indicates [1,2,3] that the majority are evolutionary related. This family currently consists of the following proteins:

- Acetylcholinesterase (EC 3.1.1.7) (AChE) [E1] from vertebrates and from *Drosophila*.
- Mammalian cholinesterase II (butyryl cholinesterase) (EC 3.1.1.8).
- 5 Acetylcholinesterase and cholinesterase II are closely related enzymes that hydrolyze choline esters [4].
- Mammalian liver microsomal carboxylesterases (EC 3.1.1.1).
- *Drosophila* esterase 6, produced in the anterior ejaculatory duct of the male insect reproductive system where it plays an important role in its
- 10 reproductive biology.
- *Drosophila* esterase P.
- *Culex pipiens* (mosquito) esterases B1 and B2.
- *Myzus persicae* (peach-potato aphid) esterases E4 and FE4.
- Mammalian bile-salt-activated lipase (BAL) [5], a multifunctional lipase
- 15 which catalyzes fat and vitamin absorption. It is activated by bile salts in infant intestine where it helps to digest milk fats.
- Insect juvenile hormone esterase (JH esterase) (EC 3.1.1.59).
- Lipases (EC 3.1.1.3) from the fungi *Geotrichum candidum* and *Candida rugosa*.
- *Caenorhabditis* gut esterase (gene *ges-1*).
- 20 - Duck fatty acyl-CoA hydrolase, medium chain (EC 3.1.2.14), an enzyme that may be associated with peroxisome proliferation and may play a role in the production of 3-hydroxy fatty acid diester pheromones.
- Membrane enclosed crystal proteins from slime mold. These proteins are, most probably esterases; the vesicles where they are found have therefore
- 25 been termed esterosomes.

So far two bacterial proteins have been found to belong to this family:

- Phenmedipham hydrolase (phenylcarbamate hydrolase), an *Arthrobacter oxidans*
- 30 plasmid-encoded enzyme (gene *pcd*) that degrades the phenylcarbamate herbicides phenmedipham and desmedipham by hydrolyzing their central carbamate linkages.
- Para-nitrobenzyl esterase from *Bacillus subtilis* (gene *pnbA*).

The following proteins, while having lost their catalytic activity, contain a domain evolutionary related to that of carboxylesterases type-B:

- 5 - Thyroglobulin (TG), a glycoprotein specific to the thyroid gland, which is the precursor of the iodinated thyroid hormones thyroxine (T4) and triiodo thyronine (T3).
- Drosophila protein neuractin (gene nrt) which may mediate or modulate cell adhesion between embryonic cells during development.
- 10 - Drosophila protein glutactin (gene glt), whose function is not known.

As is the case for lipases and serine proteases, the catalytic apparatus of esterases involves three residues (catalytic triad): a serine, a glutamate or aspartate and a histidine. The sequence around the active site serine is well
 15 conserved and can be used as a signature pattern. A conserved region located in the N-terminal section containing a cysteine involved in a disulfide bond has been selected as a second signature pattern.

-Consensus pattern: F-[GR]-G-x(4)-[LIVM SEQ ID NO:4)]-x-[LIV]-x-G-x-S-[STAG SEQ
 20 ID NO:20)]-G[S is the active site residue]

-Consensus pattern: [ED]-D-C-L-[YT]-[LIV]-[DNS]-[LIV]-[LIVFYW SEQ ID NO:147)]-x-[PQR] [C is involved in a disulfide bond]

- [1] Myers M., Richmond R.C., Oakeshott J.G. Mol. Biol. Evol. 5:113-119(1988).
- 25 [2] Krejci E., Duval N., Chatonnet A., Vincens P., Massoulie J. Proc. Natl. Acad. Sci. U.S.A. 88:6647-6651(1991).
- [3] Cygler M., Schrag J.D., Sussman J.L., Harel M., Silman I. Gentry M.K., Doctor B.P. Protein Sci. 2:366-382(1993).
- [4] Lockridge O. BioEssays 9:125-128(1988).
- 30 [5] Wang C.-S., Hartsuck J.A. Biochim. Biophys. Acta 1166:1-19(1993).

85. CPSase_L_chain (Carbamoyl-phosphate synthase (CPSase))

[1]

Medline: 94347758

Three-dimensional structure of the biotin carboxylase subunit.
of acetyl-CoA carboxylase.

5 Waldrop GL, Rayment I, Holden HM;
Biochemistry 1994;33:10249-10256.

[1]

Medline: 90285162

10 Mammalian carbamyl phosphate synthetase (CPS). DNA sequence and
evolution of the CPS domain of the Syrian hamster multifunctional
protein CAD.

Simmer JP, Kelly RE, Rinker AG Jr, Scully JL, Evans DR;
Biol Chem 1990;265:10395-10402.

15 Carbamoyl-phosphate synthase catalyzes the ATP-dependent synthesis of
carbamyl-phosphate from glutamine or ammonia and bicarbonate. This
important enzyme initiates both the urea cycle and the biosynthesis
of arginine and/or pyrimidines [2].

20 The carbamoyl-phosphate synthase (CPS) enzyme in prokaryotes is a
heterodimer of a small and large chain. The small chain promotes
the hydrolysis of glutamine to ammonia, which is used by the large
chain to synthesize carbamoyl phosphate. See CPSase_sm_chain.
The small chain has a GATase domain in the carboxyl terminus.
See GATase.

Number of members: 181

25

Carbamoyl-phosphate synthase (CPSase) catalyzes the ATP-dependent synthesis of
carbamyl-phosphate from glutamine (EC 6.3.5.5) or ammonia (EC 6.3.4.16) and
bicarbonate [1]. This important enzyme initiates both the urea cycle and the
biosynthesis of arginine and pyrimidines.

30

Glutamine-dependent CPSase (CPSase II) is involved in the biosynthesis of
pyrimidines and purines. In bacteria such as Escherichia coli, a single enzyme
is involved in both biosynthetic pathways while other bacteria have separate

enzymes. The bacterial enzymes are formed of two subunits. A small chain (gene *carA*) that provides glutamine amidotransferase activity (GATase) necessary for removal of the ammonia group from glutamine, and a large chain (gene *carB*) that provides CPSase activity. Such a structure is also present in fungi for arginine biosynthesis (genes *CPA1* and *CPA2*). In most eukaryotes, the first three steps of pyrimidine biosynthesis are catalyzed by a large multifunctional enzyme - called URA2 in yeast, rudimentary in *Drosophila* and CAD in mammals [2]. The CPSase domain is located between an N-terminal GATase domain and the C-terminal part which encompass the dihydroorotase and aspartate transcarbamylase activities.

Ammonia-dependent CPSase (CPSase I) is involved in the urea cycle in ureolytic vertebrates; it is a monofunctional protein located in the mitochondrial matrix.

The CPSase domain is typically 120 Kd in size and has arisen from the duplication of an ancestral subdomain of about 500 amino acids. Each subdomain independently binds to ATP and it is suggested that the two homologous halves act separately, one to catalyze the phosphorylation of bicarbonate to carboxy phosphate and the other that of carbamate to carbamyl phosphate.

The CPSase subdomain is also present in a single copy in the biotin-dependent enzymes acetyl-CoA carboxylase (EC 6.4.1.2) (ACC), propionyl-CoA carboxylase (EC 6.4.1.3) (PCCase), pyruvate carboxylase (EC 6.4.1.1) (PC) and urea carboxylase (EC 6.3.4.6).

Two conserved regions which are probably important for binding ATP and/or catalytic activity have been selected as signatures for the subdomain.

-Consensus pattern: [FYV]-[PS]-[LIVMC SEQ ID NO:142)]-[LIVMA SEQ ID NO:30)]-[LIVM SEQ ID NO:4)]-[KR]-[PSA]-[STA]-x(3)-[SG]-G-x-[AG]

-Consensus pattern: [LIVMF SEQ ID NO:2)]-[LIMN SEQ ID NO:148)]-E-[LIVMCA SEQ ID NO:149)]-N-[PATLIVM SEQ ID NO:150)]-[KR]-[LIVMSTAC SEQ ID NO:151)]

- [1] Simmer J.P., Kelly R.E., Rinker A.G. Jr., Scully J.L., Evans D.R.
J. Biol. Chem. 265:10395-10402(1990).
[2] Davidson J.N., Chen K.C., Jamison R.S., Musmanno L.A., Kern C.B.
5 BioEssays 15:157-164(1993).

86. CPSase_sm_chain (Carbamoyl-phosphate synthase small chain, CPSase domain)

[1]

10 Medline: 90285162

Mammalian carbamyl phosphate synthetase (CPS). DNA sequence and evolution of the CPS domain of the Syrian hamster multifunctional protein CAD.

Simmer JP, Kelly RE, Rinker AG Jr, Scully JL, Evans DR;

15 Biol Chem 1990;265:10395-10402.

The carbamoyl-phosphate synthase domain is in the amino terminus of protein.

Carbamoyl-phosphate synthase catalyzes the ATP-dependent synthesis of carbamyl-phosphate from glutamine or ammonia and bicarbonate. This
20 important enzyme initiates both the urea cycle and the biosynthesis of arginine and/or pyrimidines [1].

The carbamoyl-phosphate synthase (CPS) enzyme in prokaryotes is a heterodimer of a small and large chain. The small chain promotes the hydrolysis of glutamine to ammonia, which is used by the large
25 chain to synthesize carbamoyl phosphate. See CPSase_L_chain.

The small chain has a GATase domain in the carboxyl terminus.
See GATase.

Number of members: 46

30 Carbamoyl-phosphate synthase (CPSase) catalyzes the ATP-dependent synthesis of carbamyl-phosphate from glutamine (EC 6.3.5.5) or ammonia (EC 6.3.4.16) and bicarbonate [1]. This important enzyme initiates both the urea cycle and the biosynthesis of arginine and pyrimidines.

Glutamine-dependent CPSase (CPSase II) is involved in the biosynthesis of pyrimidines and purines. In bacteria such as *Escherichia coli*, a single enzyme is involved in both biosynthetic pathways while other bacteria have separate enzymes. The bacterial enzymes are formed of two subunits. A small chain (gene *carA*) that provides glutamine amidotransferase activity (GATase) necessary for removal of the ammonia group from glutamine, and a large chain (gene *carB*) that provides CPSase activity. Such a structure is also present in fungi for arginine biosynthesis (genes *CPA1* and *CPA2*). In most eukaryotes, the first three steps of pyrimidine biosynthesis are catalyzed by a large multifunctional enzyme - called URA2 in yeast, rudimentary in *Drosophila* and CAD in mammals [2]. The CPSase domain is located between an N-terminal GATase domain and the C-terminal part which encompass the dihydroorotase and aspartate transcarbamylase activities.

Ammonia-dependent CPSase (CPSase I) is involved in the urea cycle in ureolytic vertebrates; it is a monofunctional protein located in the mitochondrial matrix.

The CPSase domain is typically 120 Kd in size and has arisen from the duplication of an ancestral subdomain of about 500 amino acids. Each subdomain independently binds to ATP and it is suggested that the two homologous halves act separately, one to catalyze the phosphorylation of bicarbonate to carboxy phosphate and the other that of carbamate to carbamyl phosphate.

The CPSase subdomain is also present in a single copy in the biotin-dependent enzymes acetyl-CoA carboxylase (EC 6.4.1.2) (ACC), propionyl-CoA carboxylase (EC 6.4.1.3) (PCCase), pyruvate carboxylase (EC 6.4.1.1) (PC) and urea carboxylase (EC 6.3.4.6).

Two conserved regions which are probably important for binding ATP and/or catalytic activity have been selected as signatures for the subdomain.

-Consensus pattern: [FYV]-[PS]-[LIVMC SEQ ID NO:142)]-[LIVMA SEQ ID NO:30)]-[LIVM SEQ ID NO:4)]-[KR]-[PSA]-[STA]-x(3)-[SG]-G-x-[AG]

-Consensus pattern: [LIVMF SEQ ID NO:2)]-[LIMN SEQ ID NO:148)]-E-[LIVMCA SEQ ID NO:149)]-N-[PATLIVM SEQ ID NO:150)]-[KR]-[LIVMSTAC SEQ ID NO:151)]

5

[1] Simmer J.P., Kelly R.E., Rinker A.G. Jr., Scully J.L., Evans D.R.
J. Biol. Chem. 265:10395-10402(1990).

[2] Davidson J.N., Chen K.C., Jamison R.S., Musmanno L.A., Kern C.B.
BioEssays 15:157-164(1993).

10

87. CRAL_TRIO (CRAL/TRIO domain)

[1]

Medline: 98121119

15 Crystal structure of the *Saccharomyces cerevisiae* phosphatidyl-
inositol-transfer protein.

Sha B, Phillips SE, Bankaitis VA, Luo M;
Nature 1998;391:506-510.

The original profile has been extended to include the carboxyl
20 domain from the known structure of Sec14. Swiss:P10911 has not
been included in the Pfam family because it does not appear to
contain a complete structural domain.

Number of members: 39

25

88. CSD ('Cold-shock' DNA-binding domain)

[1]

Medline: 94255482

30 Crystal structure of CspA, the major cold shock
protein of *Escherichia coli*.

Schindelin H, Jiang W, Inouye M, Heinemann U;
Proc Natl Acad Sci U S A 1994;91:5119-5123.

Number of members: 121

A conserved domain of about 70 amino acids has been found in prokaryotic and eukaryotic DNA-binding proteins [1,2,3,E1]. This domain, which is known as the 'cold-shock domain' (CSD) is present in the proteins listed below.

5

- *Escherichia coli* protein CS7.4 (gene *cspA*) which is induced in response to low temperature (cold-shock protein) and which binds to and stimulates the transcription of the CCAAT-containing promoters of the *HN-S* protein and of *gyrA*.

10

- Mammalian Y box binding protein 1 (YB1). A protein that binds to the CCAAT-containing Y box of mammalian HLA class II genes.

- *Xenopus* Y box binding proteins -1 and -2 (Y1 and Y2). Proteins that bind to the CCAAT-containing Y box of *Xenopus* hsp70 genes.

- *Xenopus* B box binding protein (YB3). YB3 binds the B box promoter element of genes transcribed by RNA polymerase III.

15

- Enhancer factor I subunit A (EFI-A) (dbpB). A protein that also binds to CCAAT-motif in various gene promoters.

- DbpA, a Human DNA-binding protein of unknown specificity.

- *Bacillus subtilis* cold-shock proteins *cspB* and *cspC*.

20

- *Streptomyces clavuligerus* protein SC 7.0.

- *Escherichia coli* proteins *cspB*, *cspC*, *cspD*, *cspE* and *cspF*.

- Unr, a mammalian gene encoded upstream of the N-ras gene. Unr contains nine repeats that are similar to the CSD domain. The function of Unr is not yet known but it could be a multivalent DNA-binding protein.

25

As a signature pattern for the CSD domain, its most conserved region which is located in its N-terminal section has been selected. It must be noted that the beginning of this region is highly similar [4] to the RNP-1 RNA-binding motif.

30

-Consensus pattern: [FY]-G-F-I-x(6,7)-[DER]-[LIVM SEQ ID NO:4)]-F-x-H-x-[STKR SEQ ID NO:152)]-x-[LIVMFY SEQ ID NO:18)]

[1] Doniger J., Landsman D., Gonda M.A., Wistow G.

New Biol. 4:389-395(1992).

[2] Wistow G.

Nature 344:823-824(1990).

[3] Jones P.G., Inouye M.

5 Mol. Microbiol. 11:811-818(1994).

[4] Landsman D.

Nucleic Acids Res. 20:2861-2864(1992).

10 89. CTF_NFI (CTF/NF-I family)

Number of members: 45

Nuclear factor I (NF-I) or CCAAT box-binding transcription factor (CTF) [1,2]
(also known as TGGCA-binding proteins) are a family of vertebrate nuclear
15 proteins which recognize and bind, as dimers, the palindromic DNA sequence
5'-TGGCANNNTGCCA-3'. CTF/NF-I binding sites are present in viral and cellular
promoters and in the origin of DNA replication of Adenovirus type 2.

The CTF/NF-I proteins were first identified as nuclear factor I, a collection
20 of proteins that activate the replication of several Adenovirus serotypes
(together with NF-II and NF-III) [3]. The family of proteins was also
identified as the CTF transcription factors, before the NFI and CTF families
were found to be identical [4]. The CTF/NF-I proteins are individually capable
of activating transcription and DNA replication. The CTF/NF-I family name has
25 also been dubbed as NFI, NF-I or NF1.

In a given species, there are a large number of different CTF/NF-I proteins.
The multiplicity of CTF/NF-I is known to be generated both by alternative
splicing and by the occurrence of four different genes. The known forms of
30 NF-I genes have been classified as:

- The CTF-like factors subfamily (prototype form: CTF-1) [4]
- The NFI-X proteins.

- The NFI-A proteins.
- The NFI-B proteins.

So far, all CTF/NF-I family members appear to have similar transcription and replication activities.

CTF/NF-1 proteins contains 400 to 600 amino acids. The N-terminal 200 amino-acid sequence, almost perfectly conserved in all species and genes sequenced, mediates site-specific DNA recognition, protein dimerization and Adenovirus DNA replication. The C-terminal 100 amino acids contain the transcriptional activation domain. This activation domain is the target of gene expression regulatory pathways elicited by growth factors and it interacts with basal transcription factors and with histone H3 [6].

A perfectly conserved, highly charged 12 residue peptide located in the N-terminal part of CTF/NF-I has been selected as a specific signature for this family of proteins.

-Consensus pattern: R-K-R-K-Y-F-K-K-H-E-K-R

[1] Mermod N., O'Neill E.A., Kelly T.J., Tjian R.
Cell 58:741-753(1989).

[2] Rupp R.A.W., Kruse U., Multhaup G., Goebel U., Beyreuther K.,
Sippel A.E.
Nucleic Acids Res. 18:2607-2616(1990).

[3] Nagata K., Guggenheimer R.A., Enomoto T., Lichy J.H., Hurwitz J.
Proc. Natl. Acad. Sci. U.S.A. 79:6438-6442(1982).

[4] Santoro C., Mermod N., Andrews P.C., Tjian R.
Nature 334:2118-2224(1988).

[5] Gil G., Smith J.R., Goldstein J.L., Slaughter C.A., Orth K., Brown M.S.,
Osborne T.F.
Proc. Natl. Acad. Sci. U.S.A 85:8963-8967(1988).

[6] Alevizopoulos A., Dusserre Y., Tsai-Pflugfelder M., von der Weid T.,
Wahli W., Mermod N.

Genes Dev. 9:3051-3066(1995).

90. Calsequestrin (Calsequestrin)

5 Number of members: 13

Calsequestrin is a moderate-affinity, high-capacity calcium-binding protein of cardiac and skeletal muscle [1], where it is located in the luminal space of the sarcoplasmic reticulum terminal cisternae. Calsequestrin acts as a calcium buffer and plays an important role in the muscle excitation-contraction coupling. It is a highly acidic protein of about 400 amino acid residues that binds more than 40 moles of calcium per mole of protein. There are at least two different forms of calsequestrin: one which is expressed in cardiac muscles and another in skeletal muscles. Both forms have highly similar sequences.

Two signature sequences have been developed. The first corresponds to the N-terminus of the mature protein, the second is located just in front of the C-terminus of the protein which is composed of a highly acidic tail of variable length.

-Consensus pattern: [EQ]-[DE]-G-L-[DN]-F-P-x-Y-D-G-x-D-R-V

-Consensus pattern: [DE]-L-E-D-W-[LIVM SEQ ID NO:4)]-E-D-V-L-x-G-x-[LIVM SEQ ID NO:4)]-N-T-E-D-D-D

[1] Treves S., Vilsen B., Chiozzi P., Andersen J.P., Zorzato F.
Biochem. J. 283:767-772(1992).

30 91. Carboxyl_trans (Carboxyl transferase domain)

[1]

Medline: 93374821

Primary structure of the monomer of the 12S subunit of

transcarboxylase as deduced from DNA and characterization of the product expressed in *Escherichia coli*.

Thornton CG, Kumar GK, Haase FC, Phillips NF, Woo SB, Park VM, Magnier WJ, Shenoy BC, Wood HG, Samols D;

J Bacteriol 1993;175:5301-5308.

[2]

Medline: 93358891

Molecular evolution of biotin-dependent carboxylases.

Toh H, Kondo H, Tanabe T;

Eur J Biochem 1993;215:687-696.

All of the members in this family are biotin dependent carboxylases.

The carboxyl transferase domain carries out the following reaction;

transcarboxylation from biotin to an acceptor molecule. There are

two recognised types of carboxyl transferase. One of them uses acyl-CoA

and the other uses 2-oxo acid as the acceptor molecule of carbon dioxide.

All of the members in this family utilise acyl-CoA as the acceptor molecule.

Number of members: 47

92. Chal_stil_synt (Chalcone and stilbene synthases)

Number of members: 146

Chalcone synthases (CHS) (EC 2.3.1.74) and stilbene synthases (STS) (formerly known as resveratrol synthases) are related plant enzymes [1]. CHS is an important enzyme in flavanoid biosynthesis and STS a key enzyme in stilbene-type phytoalexin biosynthesis. Both enzymes catalyze the addition of three molecules of malonyl-CoA to a starter CoA ester (a typical example is 4-coumaroyl-CoA), producing either a chalcone (with CHS) or stilbene (with STS).

These enzymes are proteins of about 390 amino-acid residues. A conserved cysteine residue, located in the central section of these proteins, has been

shown [2] to be essential for the catalytic activity of both enzymes and probably represents the binding site for the 4-coumaryl-CoA group. The region around this active site residue is well conserved and can be used as a signature pattern.

5

In addition to the plant enzymes, this family also includes *Bacillus subtilis* bcsA.

10

-Consensus pattern: R-[LIVMFYS SEQ ID NO:153)]-x-[LIVM SEQ ID NO:4)]-x-[QHG]-x-G-C-[FYNA SEQ ID NO:154)]-[GA]-G-[GA]-[STAV SEQ ID NO:105)]-x-[LIVMF SEQ ID NO:2)]-[RA] [C is the active site residue]

[1] Schroeder J., Schroeder G.

Z. Naturforsch. 45C:1-8(1990).

15

[2] Lanz T., Tropf S., Marner F.-J., Schroeder J., Schroeder G.

J. Biol. Chem. 266:9971-9976(1991).

93. Chorismate_synt (Chorismate synthase)

20

Number of members: 19

25

Chorismate synthase (EC 4.6.1.4) catalyzes the last of the seven steps in the shikimate pathway which is used in prokaryotes, fungi and plants for the biosynthesis of aromatic amino acids. It catalyzes the 1,4-trans elimination of the phosphate group from 5-enolpyruvylshikimate-3-phosphate (EPSP) to form chorismate which can then be used in phenylalanine, tyrosine or tryptophan biosynthesis. Chorismate synthase requires the presence of a reduced flavin mononucleotide (FMNH₂ or FADH₂) for its activity.

30

Chorismate synthase from various sources shows [1,2] a high degree of sequence conservation. It is a protein of about 360 to 400 amino-acid residues. Three signature patterns have been developed from conserved regions rich in basic residues (mostly arginines). The first is in the N-terminal section, the

143

second is central and the third is C-terminal.

-Consensus pattern: G-E-S-H-[GC]-x(2)-[LIVM SEQ ID NO:4)]-[GTV]-x-[LIVM SEQ ID NO:4)](2)-[DE]-G-x-[PV]

5

-Consensus pattern: [GE]-R-[SA](2)-[SAG]-R-[EV]-[ST]-x(2)-[RH]-V-x(2)-G

-Consensus pattern: R-[SH]-D-[PSV]-[CSAV SEQ ID NO:155)]-x(4)-[GAI]-x-[IVGSP SEQ ID NO:156)]-[LIVM SEQ ID NO:4)]-x-E-[STAH SEQ ID NO:157)]-[LIVM SEQ ID NO:4)]

10 [1] Schaller A., Schmid J., Leibinger U., Amrhein N.

J. Biol. Chem. 266:21434-21438(1991).

[2] Jones D.G.L., Reusser U., Braus G.H.

Mol. Microbiol. 5:2143-2152(1991).

15

94. Clat_adaptor_s (Clathrin adaptor complex small chain)

Number of members: 21

20

Clathrin coated vesicles (CCV) mediate intracellular membrane traffic such as receptor mediated endocytosis. In addition to clathrin, the CCV are composed of a number of other components including oligomeric complexes which are known as adaptor or clathrin assembly proteins (AP) complexes [1]. The adaptor complexes are believed to interact with the cytoplasmic tails of membrane proteins, leading to their selection and concentration. In mammals two type of adaptor complexes are known: AP-1 which is associated with the Golgi complex and AP-2 which is associated with the plasma membrane. Both AP-1 and AP-2 are heterotetramers that consist of two large chains - the adaptins - (gamma and beta' in AP-1; alpha and beta in AP-2); a medium chain (AP47 in AP-1; AP50 in AP-2) and a small chain (AP19 in AP-1; AP17 in AP-2).

30

The small chains of AP-1 and AP-2 are evolutionary related proteins of about 18 Kd. Homologs of AP17 and AP19 have also been found in yeast (genes APS1/YAP19 and APS2/YAP17) [2,3,4]. AP17 and AP19 are also related to the zeta-

chain [5] of coatomer (zeta-cop), a cytosolic protein complex that reversibly associates with Golgi membranes to form vesicles that mediate biosynthetic protein transport from the endoplasmic reticulum, via the Golgi up to the trans Golgi network.

5

A conserved region in the central section of these proteins has been selected as a signature pattern.

-Consensus pattern: [LIVM SEQ ID NO:4]](2)-Y-[KR]-x(4)-L-Y-F

10

[1] Pearse B.M., Robinson M.S.

Annu. Rev. Cell Biol. 6:151-171(1990).

[2] Kirchhausen T., Davis A.C., Frucht S., O'Brine Greco B., Payne G.S.,
Tubb B.

15

J. Biol. Chem. 266:11153-11157(1991).

[3] Nakai M., Takada T., Endo T.

Biochim. Biophys. Acta 1174:282-284(1993).

[4] Phan H.L., Finlay J.A., Chu D.S., Tan P.K., Kirchhausen T., Payne G.S.
EMBO J. 13:1706-1717(1994).

20

[5] Kuge O., Hara-Kuge S., Orci L., Ravazzola M., Amherdt M., Tanigawa G.,
Wieland F.T., Rothman J.E.
J. Cell Biol. 123:1727-1734(1993).

25

95. Clathrin_lg_ch (Clathrin light chain.)

Number of members: 8

Clathrin [1,2] is the major coat-forming protein that encloses vesicles such as coated pits and forms cell surface patches involved in membrane traffic within eukaryotic cells. The clathrin coats (called triskelions) are composed of three heavy chains (180 Kd) and three light chains (23 to 27 Kd).

30

The clathrin light chains [3], which may help to properly orient the assembly

and disassembly of the clathrin coats, bind non-covalently to the heavy chain, they also bind calcium and interact with the hsc70 uncoating ATPase.

- In higher eukaryotes two genes code for distinct but related light chains:

5 LC(a) and LC(b). Each of the two genes can yield, by tissue-specific alternative splicing, two separate forms which differ by the insertion of a sequence of respectively thirty or eighteen residues. There is, in the N-terminal part of the clathrin light chains a domain of twenty one amino acid residues which is perfectly conserved in LC(a) and LC(b).

10 - In yeast there is a single light chain (gene CLC1) whose sequence is only distantly related to that of higher eukaryotes.

Two signature patterns have been developed for clathrin light chains. The first pattern is a heptapeptide from the center of the conserved N-terminal region
15 of eukaryotic light chains; the second pattern is derived from a positively charged region located in the C-terminal extremity of all known clathrin light chains.

-Consensus pattern: F-L-A-Q-Q-E-S

20

[1] Keen J.H.

Annu. Rev. Biochem. 59:415-438(1990).

[2] Brodsky F.M.

Science 242:1396-1402(1988).

25 [3] Brodsky F.M., Hill B.L., Acton S.L., Naethke I., Wong D.H.,

Ponnambalam S., Parham P.

Trends Biochem. Sci. 16:208-213(1991).

30 96. (Clathrin repeat) 7-fold repeat in Clathrin and VPS

Each repeat is about 140 amino acids long. The repeats occur in the arm region of the Clathrin heavy chain.

Number of members: 79

[1]

Medline: 92191269

Folding and trimerization of clathrin subunits at the triskelion hub.

- 5 Nathke IS, Heuser J, Lupas A, Stock J, Turck CW, Brodsky FM;
 Cell 1992;68:899-910. [2]

Medline: 88097376

Clathrin heavy chain: molecular cloning and complete primary structure.

- 10 Kirchhausen T, Harrison SC, Chow EP, Mattaliano RJ,
 Ramachandran KL, Smart J, Brosius J;
 Proc Natl Acad Sci U S A 1987;84:8805-8809.

- 15 97. Collagen (Collagen triple helix repeat (20 copies))

[1] Medline: 94059583

New members of the collagen superfamily

Mayne R, Brewton RG;

Curr Opin Cell Biol 1993;5:883-890.

- 20 Scurvy is associated with collagens.

Members of this family belong to the collagen superfamily [1].

Collagens are generally extracellular structural proteins involved in formation of connective tissue structure.

- 25 The alignment contains 20 copies of the G-X-Y repeat that
 forms a triple helix. The first position of the repeat is
 glycine, the second and third positions can be any residue
 but are frequently proline and hydroxyproline. Collagens
 are post translationally modified by proline hydroxylase
 to form the hydroxyproline residues. Defective

- 30 hydroxylation is the cause of scurvy.

Some members of the collagen superfamily are not involved in connective tissue structure but share the same triple helical structure.

Number of members: 2125

98. Coprogen_oxidas (Coproporphyrinogen III oxidase)

5 Number of members: 12

Coproporphyrinogen III oxidase (EC 1.3.3.3) (coproporphyrinogenase) [1,2] catalyzes the oxidative decarboxylation of coproporphyrinogen III into protoporphyrinogen IX, a common step in the pathway for the biosynthesis of porphyrins such as heme, chlorophyll or cobalamin.

10

Coproporphyrinogen III oxidase is an enzyme that requires iron for its activity. A cysteine seems to be important for the catalytic mechanism [3]. Sequences from a variety of eukaryotic and prokaryotic sources show that this enzyme has been evolutionarily conserved. A highly conserved region in
15 the central part of the sequence has been selected as a signature pattern. This region contains the only conserved cysteine and is rich in charged amino acids.

20

-Consensus pattern: K-x-W-C-x(2)-[FYH](3)-[LIVM SEQ ID NO:4)]-x-H-R-x-E-x-R-G-[LIVM SEQ ID NO:4)]-G-G-[LIVM SEQ ID NO:4)]-F-F-D

[1] Xu K., Elliott T.

J. Bacteriol. 175:4990-4999(1993).

[2] Kohno H., Furukawa T., Yoshinaga T., Tokunaga R., Taketani S.

25

J. Biol. Chem. 268:21359-21363(1993).

[3] Camadro J.M., Chambon H., Jolles J., Labbe P.

Eur. J. Biochem. 156:579-587(1986).

[4] Xu K., Elliott T.

J. Bacteriol. 176:3196-3203(1994).

30

99. Corona_nucleoca (Coronavirus nucleocapsid protein)

[1]

Medline: 98087828

Identification of a specific interaction between the coronavirus mouse hepatitis virus A59 nucleocapsid protein and packaging signal.

- 5 Molenkamp R, Spaan WJ;
Virology 1997;239:78-86.

Number of members: 44

- 10 100. Cu-oxidase (Multicopper oxidase)
[1]

Medline: 90126844

The blue oxidases, ascorbate oxidase, laccase and ceruloplasmin.
Modelling and structural relationships.

- 15 Messerschmidt A, Huber R;
Eur J Biochem 1990;187:341-352.

Number of members: 150

- 20 Multicopper oxidases [1,2] are enzymes that possess three spectroscopically
different copper centers. These centers are called: type 1 (or blue), type 2
(or normal) and type 3 (or coupled binuclear). The enzymes that belong to
this family are:

- 25 - Laccase (EC 1.10.3.2) (urishiol oxidase), an enzyme found in fungi and
plants, which oxidizes many different types of phenols and diamines.
- Ascorbate oxidase (EC 1.10.3.3), a higher plant enzyme.
- Ceruloplasmin (EC 1.16.3.1) (ferroxidase), a protein found in the serum of
mammals and birds, which oxidizes a great variety of inorganic and organic
substances. Structurally ceruloplasmin exhibits internal sequence homology,
30 and seem to have evolved from the triplication of a copper-binding domain
similar to that found in laccase and ascorbate oxidase.

In addition to the above enzymes there are a number of proteins which, on the

basis of sequence similarities, can be said to belong to this family. These proteins are:

- Copper resistance protein A (copA) from a plasmid in *Pseudomonas syringae*.

5 This protein seems to be involved in the resistance of the microbial host to copper.

- Blood coagulation factor V (Fa V).

- Blood coagulation factor VIII (Fa VIII) [E1].

- Yeast FET3 [3], which is required for ferrous iron uptake.

10 - Yeast hypothetical protein YFL041w and SpAC1F7.08, the fission yeast homolog.

Factors V and VIII act as cofactors in blood coagulation and are structurally similar [4]. Their sequence consists of a triplicated A domain, a B domain and
15 a duplicated C domain; in the following order: A-A-B-A-C-C. The A-type domain is related to the multicopper oxidases.

Two signature patterns have been developed for these proteins. Both patterns are derived from the same region, which in ascorbate oxidase, laccase, in the
20 third domain of ceruloplasmin, and in copA, contains five residues that are known to be involved in the binding of copper centers. The first pattern does not make any assumption on the presence of copper-binding residues and thus can detect domains that have lost the ability to bind copper (such as those in Fa V and Fa VIII), while the second pattern is specific to copper-binding
25 domains.

-Consensus pattern: G-x-[FYW]-x-[LIVMFYW SEQ ID NO:26])-x-[CST]-x(8)-G-[LM]-x(3)-[LIVMFYW SEQ ID NO:26])

-Consensus pattern: H-C-H-x(3)-H-x(3)-[AG]-[LM]

30 [The first two H's are copper type 3 binding residues]

[The C, the 3rd H, and L or M are copper type 1 ligands]

101. Cullin (Cullin family)

Number of members: 24

The following proteins are collectively termed cullins [1]:

- *Caenorhabditis elegans* cul-1 (or lin-19), a protein required for developmentally programmed transitions from the G1 phase of the cell cycle to the G0 phase or the apoptotic pathway.
- *Caenorhabditis elegans* cul-2, cul-3, cul-4 (F45E12.3), cul-5 (ZK856.1) and cul-6 (K08E7.7).
- Mammalian CUL1, CUL2, CUL3, CUL4A and CUL4B.
- Mammalian vasopressin-activated calcium-mobilizing receptor (VACM-1), a kidney-specific protein thought to form a cell surface receptor [2] but which does not have any structural hallmarks of a receptor.
- *Drosophila* lin19.
- Yeast CDC53 [3], which acts in concert with CDC4 and UBC3 (CDC34) to control the G1-to-S phase transition.
- Yeast hypothetical protein YGR003w.
- Fission yeast hypothetical protein SpAC24H6.03.

The cullins are hydrophilic proteins of 740 to 815 amino acids. The C-terminal extremity is the most conserved part of these proteins. A signature pattern has been developed from that region.

-Consensus pattern: [LIV]-K-x(2)-[LIV]-x(2)-L-I-[DEQ]-[KRHNQ SEQ ID NO:158)]-x-Y-[LIVM SEQ ID NO:4)]-x-R-x(6,7)-[FY]-x-Y-x-[SA]>

[1] Kipreos E.T., Lander L.E., Wing J.P., He W.W., Hedgecock E.M.
Cell 85:829-839(1996).

[2] Burnatowska-Hledin M.A., Spielman W.S., Smith W.L., Shi P., Meyer J.M.,
Dewitt D.L.
Am. J. Physiol. 268:f1198-F1210(1995).

[3] Mathias N., Johnson S.L., Winey M., Adams A.E., Goetsch L., Pringle J.R.,

Byers B., Goebel M.G.

Mol. Cell. Biol. 16:6634-6643(1996).

5 102. (Cu_amine_oxid)

Copper amine oxidase signatures

Amine oxidases (AO) [1] are enzymes that catalyze the oxidation of a wide range of biogenic amines including many neurotransmitters, histamine and xenobiotic amines. There are two classes of amine oxidases: flavin-containing (EC 1.4.3.4) and copper-containing (EC 1.4.3.6).

10

Copper-containing AO is found in bacteria, fungi, plants and animals, it is an homodimeric enzyme that binds one copper ion per subunit as well as a 2,4,5- trihydroxyphenylalanine quinone (or topaquinone) (TPQ) cofactor. This cofactor is derived from a tyrosine residue.

15 Two signature patterns were derived for copper AO, the first one contains the tyrosine which give rises to the TPQ cofactor while the second one contains one of the three histidines that bind the copper atom [2].

Consensus pattern[LIVM SEQ ID NO:4)]-[LIVMA SEQ ID NO:30)]-[LIVMF SEQ ID
20 NO:2)]-x(4)-[ST]-x(2)-N-Y-[DE]-[YN] [The first Y gives rises to TPQ] Sequences known to belong to this class detected by the patternALL.

Consensus patternT-x-[GS]-x(2)-H-[LIVMF SEQ ID NO:2)]-x(3)-E-[DE]-x-P [H is a copper
25 ligand] Sequences known to belong to this class detected by the pattern ALL, except for lentil AO.

[1] Knowles P.F., Dooley D.M. (In) Metal ions in biological systems; Sigel H., Sigel A., Eds., 30:361- 403, Marcel Dekker, New-York, (1993).

[2] Parsons M.R., Convery M.A., Wilmot C.M., Yadav K.D.S., Blakeley V., Corner A.S.,
30 Phillips S.E.V., McPherson M.J., Knowles P.F. Structure 3:1171-1184(1995).

103. Cys-protease (Cysteine protease)

Number of members: 358

Eukaryotic thiol proteases (EC 3.4.22.-) [1] are a family of proteolytic enzymes which contain an active site cysteine. Catalysis proceeds through a thioester intermediate and is facilitated by a nearby histidine side chain; an asparagine completes the essential catalytic triad. The proteases which are currently known to belong to this family are listed below (references are only provided for recently determined sequences).

- 10 - Vertebrate lysosomal cathepsins B (EC 3.4.22.1), H (EC 3.4.22.16), L (EC 3.4.22.15), and S (EC 3.4.22.27) [2].
- Vertebrate lysosomal dipeptidyl peptidase I (EC 3.4.14.1) (also known as cathepsin C) [2].
- Vertebrate calpains (EC 3.4.22.17). Calpains are intracellular calcium-activated thiol protease that contain both a N-terminal catalytic domain and a C-terminal calcium-binding domain.
- 15 - Mammalian cathepsin K, which seems involved in osteoclastic bone resorption [3].
- Human cathepsin O [4].
- 20 - Bleomycin hydrolase. An enzyme that catalyzes the inactivation of the antitumor drug BLM (a glycopeptide).
- Plant enzymes: barley aleurain (EC 3.4.22.16), EP-B1/B4; kidney bean EP-C1, rice bean SH-EP; kiwi fruit actinidin (EC 3.4.22.14); papaya latex papain (EC 3.4.22.2), chymopapain (EC 3.4.22.6), caricain (EC 3.4.22.30), and
- 25 proteinase IV (EC 3.4.22.25); pea turgor-responsive protein 15A; pineapple stem bromelain (EC 3.4.22.32); rape COT44; rice oryzain alpha, beta, and gamma; tomato low-temperature induced, *Arabidopsis thaliana* A494, RD19A and RD21A.
- House-dust mites allergens DerP1 and EurM1.
- 30 - Cathepsin B-like proteinases from the worms *Caenorhabditis elegans* (genes gcp-1, cpr-3, cpr-4, cpr-5 and cpr-6), *Schistosoma mansoni* (antigen SM31) and *Japonica* (antigen SJ31), *Haemonchus contortus* (genes AC-1 and AC-2), and *Ostertagia ostertagi* (CP-1 and CP-3).

- Slime mold cysteine proteinases CP1 and CP2.
- Cruzipain from *Trypanosoma cruzi* and *brucei*.
- Throphozoite cysteine proteinase (TCP) from various *Plasmodium* species.
- Proteases from *Leishmania mexicana*, *Theileria annulata* and *Theileria parva*.
- 5 - Baculoviruses cathepsin-like enzyme (v-cath).
- *Drosophila* small optic lobes protein (gene sol), a neuronal protein that contains a calpain-like domain.
- Yeast thiol protease BLH1/YCP1/LAP3.
- *Caenorhabditis elegans* hypothetical protein C06G4.2, a calpain-like
- 10 protein.

Two bacterial peptidases are also part of this family:

- Aminopeptidase C from *Lactococcus lactis* (gene pepC) [5].
- 15 - Thiol protease tpr from *Porphyromonas gingivalis*.

Three other proteins are structurally related to this family, but may have lost their proteolytic activity.

- 20 - Soybean oil body protein P34. This protein has its active site cysteine replaced by a glycine.
- Rat testin, a sertoli cell secretory protein highly similar to cathepsin L but with the active site cysteine is replaced by a serine. Rat testin should not be confused with mouse testin which is a LIM-domain protein (see
- 25 <PDOC00382>).
- *Plasmodium falciparum* serine-repeat protein (SERA), the major blood stage antigen. This protein of 111 Kd possesses a C-terminal thiol-protease-like domain [6], but the active site cysteine is replaced by a serine.
- 30 The sequences around the three active site residues are well conserved and can be used as signature patterns.

-Consensus pattern: Q-x(3)-[GE]-x-C-[YW]-x(2)-[STAGC SEQ ID NO:45)]-[STAGCV SEQ ID NO:159)] [C is the active site residue]

-Consensus pattern: [LIVMGSTAN SEQ ID NO:160)]-x-H-[GSACE SEQ ID NO:161)]-[LIVM SEQ ID NO:4)]-x-[LIVMAT SEQ ID NO:162)](2)-G-x-[GSADNH SEQ ID NO:163)] [H is the active site residue]

-Consensus pattern: [FYCH SEQ ID NO:164)]-[WI]-[LIVT SEQ ID NO:165)]-x-[KRQAG SEQ ID NO:166)]-N-[ST]-W-x(3)-[FYW]-G-x(2)-G-[LFYW SEQ ID NO:167)]-[LIVMFYG SEQ ID NO:168)]-x-[LIVMF SEQ ID NO:2)] [N is the active site residue]

[1] Dufour E. Biochimie 70:1335-1342(1988).

[2] Kirschke H., Barrett A.J., Rawlings N.D. Protein Prof. 2:1587-1643(1995).

[3] Shi G.-P., Chapman H.A., Bhairi S.M., Deleeuw C., Reddy V.Y., Weiss S.J. FEBS Lett. 357:129-134(1995).

[4] Velasco G., Ferrando A.A., Puente X.S., Sanchez L.M., Lopez-Otin C. J. Biol. Chem. 269:27136-27142(1994).

[5] Chapot-Chartier M.P., Nardi M., Chopin M.C., Chopin A., Gripon J.C. Appl. Environ. Microbiol. 59:330-333(1993).

[6] Higgins D.G., McConnell D.J., Sharp P.M. Nature 340:604-604(1989).

[7] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:461-486(1994).

104. Cys_Met_Meta_PP (Cys/Met metabolism PLP-dependent enzyme)

[1] Medline: 96428687

Crystal structure of the pyridoxal-5'-phosphate dependent cystathionine beta-lyase from Escherichia coli at 1.83 Å.

Clausen T, Huber R, Laber B, Pohlenz HD, Messerschmidt A;

J Mol Biol 1996;262:202-224.

[1] Medline: 99059720

Crystal structure of Escherichia coli cystathionine gamma-synthase at 1.5 Å resolution.

Clausen T, Huber R, Prade L, Wahl MC, Messerschmidt A;

EMBO J 1998;17:6827-6838.

Database Reference: SCOP; 1cs1; fa; [SCOP-USA][CATH-PDBSUM]

This family includes enzymes involved in cysteine and methionine metabolism. The following are members:

Cystathionine gamma-lyase,
5 Cystathionine gamma-synthase,
Cystathionine beta-lyase,
Methionine gamma-lyase,
OAH/OAS sulfhydrylase,
O-succinylhomoserine sulphhydrylase

10 All of these members participate in slightly different reactions.

All these enzymes use PLP (pyridoxal-5'-phosphate) as a cofactor.

Number of members: 52

15 A number of pyridoxal-dependent enzymes involved in the metabolism of cysteine, homocysteine and methionine have been shown [1,2] to be evolutionary related. These are:

- Cystathionine gamma-lyase (EC 4.4.1.1) (gamma-cystathionase), which catalyzes the transformation of cystathionine into cysteine, oxobutanoate and ammonia. This is the final reaction in the transsulfuration pathway that leads from methionine to cysteine in eukaryotes.
- 20 - Cystathionine gamma-synthase (EC 4.2.99.9), which catalyzes the conversion of cysteine and succinyl-homoserine into cystathionine and succinate: the first step in the biosynthesis of methionine from cysteine in bacteria (gene metB).
- 25 - Cystathionine beta-lyase (EC 4.4.1.8) (beta-cystathionase), which catalyzes the conversion of cystathionine into homocysteine, pyruvate and ammonia: the second step in the biosynthesis of methionine from cysteine in bacteria (gene metC).
- 30 - Methionine gamma-lyase (EC 4.4.1.11) (L-methioninase) which catalyzes the transformation of methionine into methanethiol, oxobutanoate and ammonia.
- OAH/OAS sulfhydrylase, which catalyzes the conversion of acetylhomoserine into homocysteine and that of acetylserine into cysteine (gene MET17 or

MET25 in yeast).

- O-succinylhomoserine sulfhydrylase (EC 4.2.99.-).
- Yeast hypothetical protein YGL184c.
- Yeast hypothetical protein YHR112c.

5

These enzymes are proteins of about 400 amino-acid residues. The pyridoxal-P group is attached to a lysine residue located in the central section of these enzymes; the sequence around this residue is highly conserved and can be used as a signature pattern to detect this class of enzymes.

10

-Consensus pattern: [DQ]-[LIVMF SEQ ID NO:2)]-x(3)-[STAGC SEQ ID NO:45)]-[STAGCI SEQ ID NO:94)]-T-K-[FYWQ SEQ ID NO:169)]-[LIVMF SEQ ID NO:2)]-x-G-[HQ]-[SGNH SEQ ID NO:170)] [K is the pyridoxal-P attachment site]

15

[1] Ono B.I., Tanaka K., Naito K., Heike C., Shinoda S., Yamamoto S., Ohmori S., Oshima T., Toh-E A.
J. Bacteriol. 174:3339-3347(1992).

[2] Barton A.B., Kaback D.B., Clark M.W., Keng T., Ouellette B.F.F., Storms R.K., Zeng B., Zhong W.W., Fortin N., Delaney S., Bussey H.

20

Yeast 9:363-369(1993).

105. Cyt_reductase

FAD/NAD-binding Cytochrome reductase

25

Number of members: 60

[1] Medline: 95111952

Crystal structure of the FAD-containing fragment of corn nitrate reductase at 2.5 Å resolution: relationship to other flavoprotein reductases.

30

Lu G, Campbell WH, Schneider G, Lindqvist Y;
Structure 1994;2:809-821.

[2] Medline: 92084635

The sequence of squash NADH:nitrate reductase and its

relationship to the sequences of other flavoprotein
oxidoreductases. A family of flavoprotein pyridine
nucleotide cytochrome reductases.

Hyde GE, Crawford NM, Campbell WH;

5 J Biol Chem 1991;266:23542-23547.

106. Cytidylyltrans

Phosphatidate cytidylyltransferase

10 Number of members: 21

Phosphatidate cytidylyltransferase (EC 2.7.7.41) [1,2,3] (also known as CDP-
diacylglycerol synthase) (CDS) is the enzyme that catalyzes the synthesis of
CDP-diacylglycerol from CTP and phosphatidate (PA). CDP-diacylglycerol is an
15 important branch point intermediate in both prokaryotic and eukaryotic
organisms. CDS is a membrane-bound enzyme. A conserved region located in the
C-terminal part has been selected as a signature pattern.

-Consensus pattern: S-x-[LIVMF SEQ ID NO:2)]-K-R-x(4)-K-D-x-[GSA]-x(2)-[LI]-[PG]-x-
20 H-G-G-[LIVM SEQ ID NO:4)]-x-D-R-[LIVMF SEQ ID NO:2)]-D

[1] Sparrow C.P., Raetz C.R.H.

J. Biol. Chem. 260:12084-12091(1985).

[2] Shen H., Heacock P.N., Clancey C.J., Dowhan W.

25 J. Biol. Chem. 271:789-795(1996).

[3] Saito S., Goto K., Tonosaki A., Kondo H.

J. Biol. Chem. 272:9503-9509(1997).

30 107. (Cytidylyltransf) Cytidylyltransferase. This family includes: Cholinephosphate
cytidylyltransferase. Glycerol-3-phosphate cytidylyltransferase.

Number of members: 64

[1] Medline: 10208837 CTP:Phosphocholine Cytidylyltransferase: Insights into Regulatory Mechanisms and Novel Functions. Clement JM, Kent C; Biochem Biophys Res Commun 1999;257:643-650.

5

108. (cNMP binding) Cyclic nucleotide-binding domain signatures and profile

10

Proteins that bind cyclic nucleotides (cAMP or cGMP) share a structural domain of about 120 residues [1-3]. The best studied of these proteins is the prokaryotic catabolite gene activator (also known as the cAMP receptor protein) (gene *crp*) where such a domain is known to be composed of three α -helices and a distinctive eight-stranded, antiparallel β -barrel structure. Such a domain is known to exist in the following proteins: - Prokaryotic catabolite gene activator protein (CAP). - cAMP- and cGMP-dependent protein kinases (cAPK and cGPK). Both types of kinases contain two tandem copies of the cyclic nucleotide-binding domain. The cAPK's are composed of two different subunits: a catalytic chain and a regulatory chain which contains both copies of the domain. The cGPK's are single chain enzymes that include the two copies of the domain in their N-terminal section. The nucleotide specificity of cAPK and cGPK is due to an amino acid in the conserved region of β -barrel 7: a threonine that is invariant in cGPK is an alanine in most cAPK. - Vertebrate cyclic nucleotide-gated ion-channels. Two such cation channels have been fully characterized. One is found in rod cells where it plays a role in visual signal transduction. It specifically binds to cGMP leading to an opening of the channel and thereby causing a depolarization of rod photoreceptors. In olfactory epithelium a similar, cAMP-binding, channel plays a role in odorant signal transduction. There are six invariant amino acids in this domain, three of which are glycine residues that are thought to be essential for

15

20

25

30

First consensus pattern: [LIVM SEQ ID NO:4)]-[VIC]-x(2)-G-[DENQTA SEQ ID NO:144)]-x-[GAC]-x(2)-[LIVMFY SEQ ID NO:18)](4)-x(2)-G

Second consensus pattern: [LIVMF SEQ ID NO:2)]-G-E-x-[GAS]-[LIVM SEQ ID NO:4)]-x(5,11)-R-[STAQ SEQ ID NO:145)]-A-x-[LIVMA SEQ ID NO:30)]-x-[STACV SEQ ID NO:146)]-

- [1] Weber I.T., Shabb J.B., Corbin J.D. Biochemistry 28:6122-6127(1989).
 [2] Kaupp U.B. Trends Neurosci. 14:150-157(1991).
 [3] Shabb J.B., Corbin J.D. J. Biol. Chem. 267:5723-5726(1992).

5

109. (cadherin)

Cadherins extracellular repeated domain signature

10 Cadherins [1,2] are a family of animal glycoproteins responsible for calcium-dependent cell-cell adhesion. Cadherins preferentially interact with themselves in a homophilic manner in connecting cells; thus acting as both receptor and ligand. A wide number of tissue-specific forms of cadherins are known:

- Epithelial (E-cadherin) (also known as uvomorulin or L-CAM) (CDH1).
- 15 - Neural (N-cadherin) (CDH2).
- Placental (P-cadherin) (CDH3).
- Retinal (R-cadherin) (CDH4).
- Vascular endothelial (VE-cadherin) (CDH5).
- Kidney (K-cadherin) (CDH6).
- 20 - Cadherin-8 (CDH8).
- Osteoblast (OB-cadherin) (CDH11).
- Brain (BR-cadherin) (CDH12).
- T-cadherin (truncated cadherin) (CDH13).
- Muscle (M-cadherin) (CDH14).
- 25 - Liver-intestine (LI-cadherin).
- EP-cadherin.

30 Structurally, cadherins are built of the following domains: a signal sequence, followed by a propeptide of about 130 residues, then an extracellular domain of around 600 residues, then a transmembrane region, and finally a C-terminal cytoplasmic domain of about 150 residues. The extracellular domain can be sub- divided into five parts: there are four repeats of about 110 residues followed by a region that contains four conserved cysteines. It is suggested that the calcium-binding region of cadherins is located in the extracellular repeats.

Cadherins are evolutionary related to the desmogleins which are component of intercellular desmosome junctions involved in the interaction of plaque proteins:

- 5 - Desmoglein 1 (desmosomal glycoprotein I).
- Desmoglein 2.
- Desmoglein 3 (Pemphigus vulgaris antigen).

10 The Drosophila fat protein [3] is a huge protein of over 5000 amino acids that contains 34 cadherin-like repeats in its extracellular domain.

15 The signature pattern that was developed for the repeated domain is located in it the C-terminal extremity which is its best conserved region. The pattern includes two conserved aspartic acid residues as well as two asparagines; these residues could be implicated in the binding of calcium.

20 Consensus pattern[LIV]-x-[LIV]-x-D-x-N-D-[NH]-x-P Sequences known to belong to this class detected by the pattern ALL. Note this pattern is found in the first, second, and fourth copies of the repeated domain. In the third copy there is a deletion of one residue after the second conserved Asp.

[1] Takeichi M. Annu. Rev. Biochem. 59:237-252(1990).

[2] Takeichi M. Trends Genet. 3:213-217(1987).

25 [3] Mahoney P.A., Weber U., Onofrechuk P., Biessmann H., Bryant P.J., Goodman C.S. Cell 67:853-868(1991).

110. Calreticulin family signatures

30 Calreticulin [1] (also known as calregulin, CRP55 or HACBP) is a high-capacitycalcium-binding protein which is present in most tissues and located at the periphery of the endoplasmic (ER) and the sarcoplamic reticulum (SR)membranes. It probably plays a role in the storage of calcium in the lumen ofthe ER and SR and it may well have other important functions. Structurally, calreticulin is a protein of about 400 amino acid residues consisting of

three domains: a) An N-terminal, probably globular, domain of about 180 amino acid residues (N-domain); b) A central domain of about 70 residues (P-domain) which contains three repeats of an acidic 17 amino acid motif. This region binds calcium with a low-capacity, but a high-affinity; c) A C-terminal domain rich in acidic residues and in lysine (C-domain). This region binds calcium with a high-capacity but a low-affinity. Calreticulin is evolutionary related to the following proteins: - *Onchocerca volvulus* antigen RAL-1. RAL-1 is highly similar to calreticulin, but possesses a C-terminal domain rich in lysine and arginine and lacks acidic residues and is therefore not expected to bind calcium in that region. - Calnexin [2]. A calcium-binding protein that interacts with newly synthesized glycoproteins in the endoplasmic reticulum. It seems to play a major role in the quality control apparatus of the ER by the retention of incorrectly folded proteins. - Calmegin [3] (or calnexin-T), a testis-specific calcium-binding protein highly similar to calnexin. Three signature patterns have been developed for this family of proteins. The first two patterns are based on conserved regions in the N-domain; the third pattern corresponds to positions 4 to 16 of the repeated motif in the P-domain.

Consensus pattern: [KRHN SEQ ID NO:171)]-x-[DEQN SEQ ID NO:172)]-[DEQNK SEQ ID NO:173)]-x(3)-C-G-G-[AG]-[FY]-[LIVM SEQ ID NO:4)]-[KN]- [LIVMFY SEQ ID NO:18)](2)-

Consensus pattern: [LIVM SEQ ID NO:4)](2)-F-G-P-D-x-C-[AG]-

Consensus pattern: [IV]-x-D-x-[DENST SEQ ID NO:174)]-x(2)-K-P-[DEH]-D-W-[DEN]-

[1] Michalak M., Milner R.E., Burns K., Opas M. *Biochem. J.* 285:681-692(1992).

[2] Bergeron J.J.M., Brenner M.B., Thomas D.Y., Williams D.B. *Trends Biochem. Sci.* 19:124-128(1994).

[3] Watanabe D., Yamada K., Nishina Y., Tajima Y., Koshimizu U., Nagata A., Nishimune Y. *J. Biol. Chem.* 269:7744-7749(1994).

111. Eukaryotic-type carbonic anhydrases signature (carb_anhydrase)

Carbonic anhydrases (EC 4.2.1.1) (CA) [1,2,3,4] are zinc metalloenzymes which catalyze the reversible hydration of carbon dioxide. Eight enzymatic and evolutionary related forms of carbonic anhydrase are currently known to exist in vertebrates: three cytosolic isozymes (CA-I, CA-II and CA-III); two membrane-bound forms (CA-IV and CA-VII); a mitochondrial

form (CA-V); a secreted salivary form (CA-VI); and a yet uncharacterized isozyme [5]. In the alga *Chlamydomonas reinhardtii*, two CA isozymes have been sequenced [6]. They are periplasmic glycoproteins evolutionary related to vertebrate CAs. Some bacteria, such as *Neisseria gonorrhoeae* [7] also have a eukaryotic-type CA. CAs contain a single zinc atom bound to three conserved histidine residues. As a signature for CAs, a pattern has been developed which includes one of these zinc-binding histidines. Protein D8 from *Vaccinia* and other poxviruses is related to CAs but has lost two of the zinc-binding histidines as well as many otherwise conserved residues. This is also true of the N-terminal extracellular domain of some receptor-type tyrosine-protein phosphatases (see <PDOC00323>).

Consensus pattern: S-E-[HN]-x-[LIVM SEQ ID NO:4)]-x(4)-[FYH]-x(2)-E-[LIVMGA SEQ ID NO:175)]-H-[LIVMFA SEQ ID NO:81)](2) [The second H is a zinc ligand]-

Note: most prokaryotic CA's as well as plant chloroplast CA's belong to another, evolutionary distinct family of proteins (see <PDOC00586

[1] Deutsch H.F. Int. J. Biochem. 19:101-113(1987).

[2] Fernley R.T. Trends Biochem. Sci. 13:356-359(1988).

[3] Tashian R.E. BioEssays 10:186-192(1989).

[4] Edwards Y. Biochem. Soc. Trans. 18:171-175(1990).

[5] Skaggs L.A., Bergenhem N.C.H., Venta P.J., Tashian R.E. Gene 126:291-292(1993).

[6] Fujiwara S., Fukuzawa H., Tachiki A., Miyachi S. Proc. Natl. Acad. Sci. U.S.A. 87:9779-9783(1990).

[7] Huang S., Xue Y., Sauer-Eriksson E., Chirica L., Lindskog S., Jonsson B.H. 2.3.CO;2-"J. Mol. Biol. 283:301-310(1998).

112. Caseins alpha/beta signature

Caseins [1] are the major protein constituent of milk. Caseins can be classified into two families; the first consists of the kappa-caseins, and the second groups the alpha-s1, alpha-s2, and beta-caseins. The alpha/beta caseins are a rapidly diverging family of proteins. However two regions are conserved: a cluster of phosphorylated serine residues and the signal sequence. The signature pattern has been developed for this family of proteins based upon the last eight residues of the signal sequence.

Consensus pattern: C-L-[LV]-A-x-A-[LVF]-A -

[1] Holt C., Sawyer L. Protein Eng. 2:251-259(1988).

5 113. Catalase signatures

Catalase (EC 1.11.1.6) [1,2,3] is an enzyme, present in all aerobic cells, that decomposes hydrogen peroxide to molecular oxygen and water. Its main function is to protect cells from the toxic effects of hydrogen peroxide. In eukaryotic organisms and in some prokaryotes catalase is a molecule composed of four identical subunits. Each of the subunits binds one
10 protoheme IX group. A conserved tyrosine serves as the heme proximal side ligand. The region around this residue has been used as a first signature pattern; it also includes a conserved arginine that participates in heme-binding. A conserved histidine has been shown to be important for the catalytic mechanism of the enzyme. The region around this residue has been selected as a second signature pattern.-

15 Consensus pattern: R-[LIVMFSTAN SEQ ID NO:176)]-F-[GASTNP SEQ ID NO:177)]-Y-x-D-[AST]-[QEH] [Y is the proximal heme-binding ligand]

Consensus pattern: [IF]-x-[RH]-x(4)-[EQ]-R-x(2)-H-x(2)-[GAS]-[GASTF SEQ ID NO:178)]-[GAST SEQ ID NO:179)] [H is an active site residue]

Note: some prokaryotic catalases belong to the peroxidase family (see <[PDOC00394](#)>).

20

[1] Murthy M.R.N., Reid T.J. III, Sicignano A., Tanaka N., Rossmann M.G. J. Mol. Biol. 152:465-499(1981).

[2] Melik-Adamyany W.R., Barynin V.V., Vagin A.A., Borisov V.V., Vainshtein B.K., Fita I., Murthy M.R.N., Rossmann M.G. J. Mol. Biol. 188:63-72(1986).

25 [3] von Ossowski I., Hausner G., Loewen P.C. J. Mol. Evol. 37:71-76(1993).

114. (chitin binding) Chitin recognition or binding domain signature

A conserved domain of 43 amino acids is found in several plant and fungal proteins that have
30 a common binding specificity for oligosaccharides of N-acetylglucosamine [1]. This domain may be involved in the recognition or binding of chitin subunits. It has been found in the proteins listed below. - A number of non-leguminous plant lectins. The best characterized of these lectins are the three highly homologous wheat germ agglutinins (WGA-1, 2 and 3).

WGA is an N-acetylglucosamine/N-acetylneuraminic acid binding lectin which structurally consists of a fourfold repetition of the 43 amino acid domain. The same type of structure is found in a barley root-specific lectin as well as a rice lectin. - Plants endochitinases (EC 3.2.1.14) from class IA (see <PDOC00620>). Endochitinases are enzymes that catalyze the hydrolysis of the beta-1,4 linkages of N-acetyl glucosamine polymers of chitin. Plant chitinases function as a defense against chitin containing fungal pathogens. Class IA chitinases generally contain one copy of the chitin-binding domain at their N-terminal extremity. An exception is agglutinin/chitinase [2] from the stinging nettle *Urtica dioica* which contains two copies of the domain. - Hevein [5], a wound-induced protein found in the latex of rubber trees. - Win1 and win2, two wound-induced proteins from potato. - *Kluyveromyces lactis* killer toxin alpha subunit [3]. The toxin encoded by the linear plasmid pGKL1 is composed of three subunits: alpha, beta, and gamma. The gamma subunit harbors toxin activity and inhibits growth of sensitive yeast strains in the G1 phase of the cell cycle; the alpha subunit, which is proteolytically processed from a larger precursor that also contains the beta subunit, is a chitinase (see <PDOC00839>). In chitinases, as well as in the potato wound-induced proteins, the 43-residue domain directly follows the signal sequence and is therefore at the N-terminal of the mature protein; in the killer toxin alpha subunit it is located in the central section of the protein. The domain contains eight conserved cysteine residues which have all been shown, in WGA, to be involved in disulfide bonds. The topological arrangement of the four disulfide bonds is shown in the following figure: +-----

```

-----+ +----|-----+ |||| xxCGxxxxxxxxCxxxxCCsxxgxCgxxxxxCxxxCxxxxC |
*****|***** |||| +----+ +-----+'C': conserved cysteine involved in a
disulfide bond.'*': position of the pattern.

```

-Consensus pattern: C-x(4,5)-C-C-S-x(2)-G-x-C-G-x(4)-[FYW]-C [The five C's are involved in disulfide bonds]

[1] Wright H.T., Sandrasegaram G., Wright C.S. J. Mol. Evol. 33:283-294(1991).

[2] Lerner D.R., Raikhel N.V. J. Biol. Chem. 267:11085-11091(1992).

[3] Butler A.R., O'Donnel R.W., Martin V.J., Gooday G.W., Stark M.J.R. Eur. J. Biochem. 199:483-488(1991).

165

115. (Chitinase 1) Chitinases family 19 signatures

Chitinases (EC 3.2.1.14) [1] are enzymes that catalyze the hydrolysis of the beta-1,4-N-acetyl-D-glucosamine linkages in chitin polymers. From the viewpoint of sequence similarity chitinases belong to either family 18 or 19 in the classification of glycosyl hydrolases [2,E1].

Chitinases of family 19 (also known as classes IA or I and IB or II) are enzymes from plants that function in the defense against fungal and insect pathogens by destroying their chitin-containing cell wall. Class IA/I and IB/II enzymes differ in the presence (IA/I) or absence (IB/II) of a N-terminal chitin-binding domain (see the relevant entry <PDOC00025>). The catalytic domain of these enzymes consists of about 220 to 230 amino acid residues. Two highly conserved regions have been selected as signature patterns, the first one is located in the N-terminal section and contains one of the six cysteines which are conserved in most, if not all, of these chitinases and which is probably involved in a disulfide bond.

Consensus pattern: C-x(4,5)-F-Y-[ST]-x(3)-[FY]-[LIVMF SEQ ID NO:2)]-x-A-x(3)-[YF]-x(2)-F- [GSA]

Consensus pattern: [LIVM SEQ ID NO:4)]-[GSA]-F-x-[STAG SEQ ID NO:20)](2)-[LIVMFY SEQ ID NO:18)]-W-[FY]-W-[LIVM SEQ ID NO:4)]

[1] Flach J., Pilet P.-E., Jolles P. *Experientia* 48:701-716(1992).

[2] Henrissat B. *Biochem. J.* 280:309-316(1991).

116. chloroa_b-bind

Chlorophyll A-B binding proteins. Number of members: 211

117. chromo

The 'chromo' (CHRromatin Organization MOdifier) domain [1 to 4] is a conserved region of about 60 amino acids which was originally found in *Drosophila* modifiers of variegation, which are proteins that modify the structure of chromatin to the condensed morphology of heterochromatin, a cytologically visible condition where gene expression is repressed. In protein Polycomb, the chromo domain has been shown to be important for chromatin targeting. Proteins

that contains a chromo domain seem to fall into three classes:

- a) Proteins which have a N-terminal chromo domain followed by a region which is related to but distinct from the chromo domain and which has been termed [3] the 'chromo shadow' domain.
- b) Proteins with a single chromo domain.
- c) Proteins with paired tandem chromo domains.

Currently, this domain has been found in the following proteins:

Class A.

- Drosophila heterochromatin protein Su(var)205 (HP1).
- Human heterochromatin protein HP1 alpha.
- Mammalian modifier 1 and modifier 2.
- Fission yeast swi6, a protein involved in the repression of the silent mating-type loci mat2 and mat3.

Class B.

- Drosophila protein Polycomb (Pc).
- Mammalian modifier 3, a homolog of Pc.
- Drosophila protein Su(var)3-9, a suppressor of position-effect variegation.
- Human Mi-2 autoantigen, characteristic of dermatomyositis.
- Fungal retrotransposon polyproteins: 'skippy' from *Fusarium oxysporum*, 'grasshopper' and 'MAGGY' from *Magnaporthe grisea* and CfT-1 from *Cladosporium fulvum*.
- Fission yeast hypothetical protein SpAC18G6.02c.
- *Caenorhabditis elegans* hypothetical protein C29H12.5
- *Caenorhabditis elegans* hypothetical protein ZK1236.2.
- *Caenorhabditis elegans* hypothetical protein T09A5.8.

Class C.

- Mammalian DNA-binding/helicase proteins CHD-1 to CHD-4.
- Yeast protein CHD1.

The signature pattern for this domain corresponds to its best conserved section, which is located in its central part.

5 -Consensus pattern: [FYL]-x-[LIVMC SEQ ID NO:142)]-[KR]-W-x-[GDNR SEQ ID NO:180)]-[FYWLME SEQ ID NO:181)]-x(5,6)-[ST]-W-[ESV]-[PSTDEN SEQ ID NO:182)]-x(2,3)-[LIVMC SEQ ID NO:142)]

[1] Paro R. Trends Genet. 6:416-421(1990).

10 [2] Singh P.B., Miller J.R., Pearce J., Kothary R., Burton R.D., Paro R., James T.C., Gaunt S.J. Nucleic Acids Res. 19:789-794(1991).

[3] Aasland R., Stewart A.F. Nucleic Acids Res. 23:3168-3173(1995).

[4] Koonin E.V., Zhou S., Lucchesi J.C. Nucleic Acids Res. 23:4229-4233(1995).

15

118. citrate_synt

Citrate synthase (EC 4.1.3.7) (CS) is the tricarboxylic acid cycle enzyme that catalyzes the synthesis of citrate from oxaloacetate and acetyl-CoA in an aldol condensation. CS can directly form a carbon-carbon bond in the absence
20 of metal ion cofactors.

In prokaryotes, citrate synthase is composed of six identical subunits. In eukaryotes, there are two isozymes of citrate synthase: one is found in the mitochondrial matrix, the second is cytoplasmic. Both seem to be dimers of
25 identical chains.

There are a number of regions of sequence similarity between prokaryotic and eukaryotic citrate synthases. One of the best conserved contains a histidine which is one of three residues shown [1] to be involved in the catalytic
30 mechanism of the vertebrate mitochondrial enzyme. This region has been used as a signature pattern.

-Consensus pattern: G-[FYA]-[GA]-H-x-[IV]-x(1,2)-[RKT]-x(2)-D-[PS]-R [H is an active site residue]

[1] Karpusas M., Branchaud B., Remington S.J. Biochemistry 29:2213-2219(1990).

5

119. clpA_B

Chaperonin clpA/B

CAUTION! This family is a subfamily of the AAA superfamily. The threshold has been set very high to stop overlaps with the AAA superfamily. This entry will be subsumed by AAA in the future.

Number of members: 39

15 A number of ATP-binding proteins that are are thought to protect cells from extreme stress by controlling the aggregation of denaturation of vital cellular structures have been shown [1,2] to be evolutionary related. These proteins are listed below.

- 20 - Escherichia coli clpA, which acts as the regulatory subunit of the ATP-dependent protease clp.
- Rhodopseudomonas blastica clpA homolog.
- Escherichia coli heat shock protein clpB and homologs in other bacteria.
- Bacillus subtilis protein mecB.
- 25 - Yeast heat shock protein 104 (gene HSP104), which is vital for tolerance to heat, ethanol and other stresses.
- Neurospora heat shock protein hsp98.
- Yeast mitochondrial heat shock protein 78 (gene HSP78) [3].
- CD4A and CD4b, two highly related tomato proteins that seem to be located
- 30 in the chloroplast.
- Trypanosoma brucei protein clp.
- Porphyra purpurea chloroplast encoded clpC.

The size of these proteins range from 84 Kd (clpA) to slightly more than 100 Kd (HSP104). They all share two conserved regions of about 200 amino acids that each contains an ATP-binding site. In addition to the ATP-binding A and B motifs there are many parts in these two domains that are also conserved. Two of these regions have been selected as signature patterns. The first signature is located in the first domain, some ten residues to the C-terminal of the ATP-binding B motif. The second pattern is located in the second domain in-between the ATP-binding A and B motifs.

- 10 -Consensus pattern: D-[AI]-[SGA]-N-[LIVMF SEQ ID NO:2)](2)-K-[PT]-x-L-x(2)-G
 -Consensus pattern: R-[LIVMFY SEQ ID NO:18)]-D-x-S-E-[LIVMFY SEQ ID NO:18)]-x-E-[KRQ]-x-[STA]-x-[STA]-[KR]-[LIVM SEQ ID NO:4)]-x-G-[STA]

- [1] Gottesman S., Squires C., Pichersky E., Carrington M., Hobbs M., Mattick J.S., Dalrymple B., Kuramitsu H., Shiroza T., Foster T., Clark W.P., Ross B., Squires C.L., Maurizi M.R. Proc. Natl. Acad. Sci. U.S.A. 87:3513-3517(1990).
 [2] Parsell D.A., Sanchez Y., Stitzel J.D., Lindquist S. Nature 353:270-273(1991).
 [3] Leonhardt S.A., Fearon K., Danese P.N., Mason T.L. Mol. Cell. Biol. 13:6304-6313(1993).

20

120. cofilin_ADF

Cofilin/tropomyosin-type actin-binding proteins

[1]

- 25 Medline: 97290449

Structure determination of yeast cofilin.

Fedorov AA, Lappalainen P, Fedorov EV, Drubin DG, Almo SC;
 Nat Struct Biol 1997;4:366-369.

[2]

- 30 Medline: 97290450

Crystal structure of the actin-binding protein actophorin
 from Acanthamoeba.

Leonard SA, Gittis AG, Petrella EC, Pollard TD, Lattman EE;

Nat Struct Biol 1997;4:369-373.

[3]

Medline: 97420794

F-actin and G-actin binding are uncoupled by mutation of
conserved tyrosine residues in maize actin depolymerizing
factor.

Jiang CJ, Weeds AG, Khan S, Hussey PJ;

Proc Natl Acad Sci U S A 1997;94:9973-9978.

[4]

Medline: 97357155

Cofilin promotes rapid actin filament turnover in vivo.

Lappalainen P, Drubin DG;

Nature 1997;388:78-82.

Severs actin filaments and binds to actin monomers.

Number of members: 44

Actin-depolymerizing proteins sever actin filaments (F-actin) and/or bind to
actin monomers, or G-actin, thus preventing actin-polymerization by
sequestering the monomers. The following proteins are evolutionary related
and belong to a family of low molecular weight (137 to 166 residues) actin-
depolymerizing proteins [1,2,3,4]:

- Cofilin from vertebrates, slime mold and yeast. Cofilin binds to F-actin
and acts as a pH-dependent actin-depolymerizing protein.

- Destrin from vertebrates. Destrin binds to G-actin in a pH-independent
manner and prevents polymerization.

- *Caenorhabditis elegans* unc-60.

- *Acanthamoeba castellanii* actophorin.

- Plants actin depolymerizing factor (ADF).

The most conserved region of these proteins is a twenty amino-acid segment
that ends some 30 residues from their C-terminal extremity. This segment has
been shown [5] to be important for actin-binding.

-Consensus pattern: P-[DE]-x-[SA]-x-[LIVMT SEQ ID NO:1)]-[KR]-x-[KR]-M-[LIVM SEQ ID NO:4)]-[YA]-[STA](3)-x(3)-[LIVMF SEQ ID NO:2)]-[KR]

- 5 [1] Hawkins M., Pope B., MacIver S.K., Weeds A.G. *Biochemistry* 32:9985-9993(1993).
 [2] Iida K., Moriyama K., Matsumoto S., Kawasaki H., Nishida E., Yahara I. *Gene* 124:115-120(1993).
 [3] Quirk S., MacIver S.K., Ampe C., Doberstein S.K., Kaiser D.A., van Damme J., Vandekerckhove J., Pollard T.D. *Biochemistry* 32:8525-8533(1993).
 10 [4] McKim K.S., Matheson C., Marra M.A., Wakarchuk M.F., Baillie D.L. *Mol. Gen. Genet.* 242:346-357(1994).
 [5] Moriyama K., Yonezawa N., Sakai H., Yahara I., Nishida E. *J. Biol. Chem.* 267:7240-7244(1992).

15

121. (Complex 24kd) Respiratory-chain NADH dehydrogenase 24 Kd subunit signature
 Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complex I or
 NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex located in the inner
 mitochondrial membrane which also seems to exist in the chloroplast and in cyanobacteria (as
 20 a NADH-plastoquinone oxidoreductase). Among the 25 to 30 polypeptide subunits of this
 bioenergetic enzyme complex there is one with a molecular weight of 24 Kd (in mammals),
 which is a component of the iron-sulfur (IP) fragment of the enzyme. It seems to bind a 2Fe-
 2S iron-sulfur cluster. The 24 Kd subunit is nuclear encoded, as a precursor form with a
 transit peptide in mammals, and in *Neurospora crassa*. The 24 Kd subunit is highly similar to
 25 [3,4]: - Subunit E of *Escherichia coli* NADH-ubiquinone oxidoreductase (gene *nuoE*). -
 Subunit NQO2 of *Paracoccus denitrificans* NADH-ubiquinone oxidoreductase. A highly
 conserved region, located in the central section of this subunit containing two conserved
 cysteines that are probably involved in the binding of the 2Fe-2S center has been selected as a
 signature pattern.

30

-Consensus pattern: D-x(2)-F-[ST]-x(5)-C-L-G-x-C-x(2) [GA]-P [The two C's are putative
 2Fe-2S ligands]

- [1] Ragan C.I. *Curr. Top. Bioenerg.* 15:1-36(1987).

- [2] Weiss H., Friedrich T., Hofhaus G., Preis D. Eur. J. Biochem. 197:563-576(1991).
[3] Fearnley I.M., Walker J.E. Biochim. Biophys. Acta 1140:105-134(1992).
[4] Weidner U., Geier S., Ptock A., Friedrich T., Leif H., Weiss H. J. Mol. Biol. 233:109-122(1993).

5

122. copper-bind

Copper binding proteins, plastocyanin/azurin family

Number of members: 70

10

Blue or 'type-1' copper proteins are small proteins which bind a single copper atom and which are characterized by an intense electronic absorption band near 600 nm [1,2]. The most well known members of this class of proteins are the plant chloroplastic plastocyanins, which exchange electrons with cytochrome c6, and the distantly related bacterial azurins, which exchange electrons with cytochrome c551. This family of proteins also includes all the proteins listed below (references are only provided for recently determined sequences).

15

20

- Amicyanin from bacteria such as *Methylobacterium extorquens* or *Thiobacillus versutus* that can grow on methylamine. Amicyanin appears to be an electron receptor for methylamine dehydrogenase.

- Auracyanins A and B from *Chloroflexus aurantiacus* [3]. These proteins can donate electrons to cytochrome c-554.

25

- Blue copper protein from *Alcaligenes faecalis*.

- Cupredoxin (CPC) from cucumber peelings [4].

- Cusacyanin (basic blue protein; plantacyanin, CBP) from cucumber.

- Halocyanin from *Natrobacterium pharaonis* [5], a membrane associated copper-binding protein.

30

- Pseudoazurin from *Pseudomonas*.

- Rusticyanin from *Thiobacillus ferrooxidans*. Rusticyanin is an electron carrier from cytochrome c-552 to the a-type oxidase [6].

- Stellacyanin from the Japanese lacquer tree.

- Umecyanin from horseradish roots.

- Allergen Ra3 from ragweed. This pollen protein is evolutionary related to the above proteins, but seems to have lost the ability to bind copper.

5

Although there is an appreciable amount of divergence in the sequence of all these proteins, the copper ligand sites are conserved and a pattern which includes two of the ligands (a cysteine and a histidine) has been developed.

10

-Consensus pattern: [GA]-x(0,2)-[YSA]-x(0,1)-[VFY]-x-C-x(1,2)-[PG]-x(0,1)-H-x(2,4)-[MQ] [C and H are copper ligands]

[1] Garret T.P.J., Clingeffer D.J., Guss J.M., Rogers S.J., Freeman H.C. J. Biol. Chem. 259:2822-2825(1984).

15

[2] Ryden L.G., Hunt L.T. J. Mol. Evol. 36:41-66(1993).

[3] McManus J.D., Brune D.C., Han J., Sanders-Loehr J., Meyer T.E., Cusanovich M.A., Tollin G., Blankenship R.E. J. Biol. Chem. 267:6531-6540(1992).

[4] Mann K., Schaefer W., Thoenes U., Messerschmidt A., Mehrabian Z., Nalbandyan R. FEBS Lett. 314:220-223(1992).

20

[5] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol. Chem. 269:14939-14945(1994).

[6] Yano T., Fukumori Y., Yamanaka T. FEBS Lett. 288:159-162(1991).

25

123. Chaperonins cpn10 signature

Chaperonins [1,2] are proteins involved in the folding of proteins or the assembly of oligomeric protein complexes. They seem to assist other polypeptides in maintaining or assuming conformations which permit their correct assembly into oligomeric structures. They are found in abundance in prokaryotes, chloroplasts and mitochondria. Chaperonins form

30

oligomeric complexes and are composed of two different types of subunits: a 60 Kd protein, known as cpn60 (groEL in bacteria) and a 10 Kd protein, known as cpn10 (groES in bacteria). The cpn10 protein binds to cpn60 in the presence of MgATP and suppresses the ATPase activity of the latter. Cpn10 is a protein of about 100 amino acid residues whose

sequence is well conserved in bacteria, vertebrate mitochondria and plants chloroplast [3,4]. Cpn10 assembles as an heptamer that forms a dome[5]. As a signature pattern for cpn10, a region located in the N-terminal section of the protein was selected.

- 5 Consensus pattern: [LIVMFY SEQ ID NO:18)]-x-P-[ILT]-x-[DEN]-[KR]-[LIVMFA SEQ ID NO:81)](3)-[KREQ SEQ ID NO:78)]-x(8,9)- [SG]-x-[LIVMFY SEQ ID NO:18)](3)-

Note: this pattern is found twice in the plant chloroplast protein which consist of the tandem repeat of a cpn10 domain

- 10 [1] Ellis R.J., van der Vies S.M. Annu. Rev. Biochem. 60:321-347(1991).
 [2] Zeilstra-Ryalls J., Fayet O., Georgopoulos C. Annu. Rev. Microbiol. 45:301-325(1991).
 [3] Hartman D.J., Hoogenraad N.J., Condrón R., Hoj P.B. Proc. Natl. Acad. Sci. U.S.A. 89:3394-3398(1992).
 [4] Bertsch U., Soll J., Seetharam R., Viitanen P.V. Proc. Natl. Acad. Sci. U.S.A. 89:8696-
 15 8700(1992).
 [5] Hunt J.F., Weaver A.J., Landry S.J., Gierasch L., Deisenhofer J. Nature 379:37-45(1996).

124. Chaperonins cpn60 signature (cpn60_TCP1)

- 20 Chaperonins [1,2] are proteins involved in the folding of proteins or the assembly of oligomeric protein complexes. Their role seems to be to assist other polypeptides to maintain or assume conformations which permit their correct assembly into oligomeric structures. They are found in abundance in prokaryotes, chloroplasts and mitochondria. Chaperonins form oligomeric complexes and are composed of two different types of subunits: a 60 Kd
 25 protein, known as cpn60 (groEL in bacteria) and a 10 Kd protein, known as cpn10 (groES in bacteria). The cpn60 protein shows weak ATPase activity and is a highly conserved protein of about 550 to 580 amino acid residues which has been described by different names in different species: - Escherichia coli groEL protein, which is essential for the growth of the bacteria and the assembly of several bacteriophages. - Cyanobacterial groEL analogues. -
 30 Mycobacterium tuberculosis and leprae 65 Kd antigen, Coxiella burnetii heat shock protein B (gene htpB), Rickettsia tsutsugamushi major antigen 58, and Chlamydial 57 Kd hypersensitivity antigen (gene hypB). - Chloroplast RuBisCO subunit binding-protein alpha and beta chains, which bind ribulose biphosphate carboxylase small and large subunits and

are implicated in the assembly of the enzyme oligomer. - Mammalian mitochondrial matrix protein P1 (mitonin or P60). - Yeast HSP60 protein, a mitochondrial assembly factor. As a signature pattern for these proteins, a rather well-conserved region of twelve residues, located in the last third of the cpn60 sequence was chosen.

Consensus pattern: A-[AS]-x-[DEQ]-E-x(4)-G-G-[GA]-

[1] Ellis R.J., van der Vies S.M. Annu. Rev. Biochem. 60:321-347(1991).

[2] Zeilstra-Ryalls J., Fayet O., Georgopoulos C. Annu. Rev. Microbiol. 45:301-325(1991).

Chaperonins TCP-1 signatures (cpn60_TCP1)

The TCP-1 protein [1,2] (Tailless Complex Polypeptide 1) was first identified in mice where it is especially abundant in testis but present in all cell types. It has since been found and characterized in many other mammalian species, in *Drosophila* and in yeast. TCP-1 is a highly conserved protein of about 60 Kd (556 to 560 residues) which participates in a hetero-oligomeric 900 Kd double-torus shaped particle [3] with 6 to 8 other different subunits. These subunits, the chaperonin containing TCP-1 (CCT) subunit beta, gamma, delta, epsilon, zeta and eta are evolutionary related to TCP-1 itself [4,5]. The CCT is known to act as a molecular chaperone for tubulin, actin and probably some other proteins. The CCT subunits are highly related to archaebacterial counterparts: - TF55 and TF56 [6], a molecular chaperone from *Sulfolobus shibatae*. TF55 has ATPase activity, is known to bind unfolded polypeptides and forms a oligomeric complex of two stacked nine-membered rings. - Thermosome [7], from *Thermoplasma acidophilum*. The thermosome is composed of two subunits (alpha and beta) and also seems to be a chaperone with ATPase activity. It forms an oligomeric complex of eight-membered rings. The TCP-1 family of proteins are weakly, but significantly [8], related to the cpn60/groEL chaperonin family (see <PDOC00268>). As signature patterns of this family of chaperonins, three conserved regions located in the N-terminal domain were chosen.

Consensus pattern: [RKEL SEQ ID NO:183)]-[ST]-x-[LMFY SEQ ID NO:184)]-G-P-x-[GSA]-x-x-K-[LIVMF SEQ ID NO:2)](2)-

176

Consensus pattern: [LIVM SEQ ID NO:4)]-[TS]-[NK]-D-[GA]-[AVNHK SEQ ID NO:185)]-[TAV]-[LIVM SEQ ID NO:4)](2)-x(2)- [LIVM SEQ ID NO:4)]-x-[LIVM SEQ ID NO:4)]-x-[SNH]-[PQH]-

Consensus pattern: Q-[DEK]-x-x-[LIVMGTA SEQ ID NO:186)]-[GA]-D-G-T-

5

[1] Ellis J. Nature 358:191-192(1992).

[2] Nelson R.J., Craig E.A. Curr. Biol. 2:487-489(1992).

[3] Lewis V.A., Hynes G.M., Zheng D., Saibil H., Willison K.R. Nature 358:249-252(1992).

[4] Kubota H., Hynes G., Carne A., Ashworth A., Willison K.R. Curr. Biol. 4:89-99(1994)

10 [5] Kim S., Willison K.R., Horwich A.L. Trends Biochem. Sci. 20:543-548(1994).

[6] Trent J.D., Nimmesgern E., Wall J.S., Hartl F.U., Horwich A.L. Nature 354:490-493(1991).

[7] Waldmann T., Lupas A., Kellermann J., Peters J., Baumeister W. Biol. Chem. Hoppe-Seyler 376:119-126(1995).

15 [8] Hemmingsen S.M. Nature 357:650-650(1992).

125. cyclin (Cyclins)

The cyclins include an internal duplication, which is related to that found in TFIIB and the RB protein.

20

[1]

Medline: 94203808

Evidence for a protein domain superfamily shared by the cyclins, TFIIB and RB/p107.

25 Gibson TJ, Thompson JD, Blocker A, Kouzarides T;

Nucleic Acids Res 1994;22:946-952.

[2]

Medline: 96164440

The crystal structure of cyclin A

30 Brown NR, Noble MEM, Endicott JA, Garman EF, Wakatsuki S,

Mitchell E, Rasmussen B, Hunt T, Johnson LN;

Structure. 1995;3:1235-1247.

Complex of cyclin and cyclin dependant kinase.

[3]

Medline: 96313126

Structural basis of cyclin-dependant kinase activation by phosphorylation.

5 Russo AA, Jeffrey PD, Pavletich NP;
Nat Struct Biol. 1996;3:696-700.

Cyclins regulate cyclin dependant kinases (CDKs).

The most divergent prosite members have been included. Swiss:P22674
the Uracil-DNA glycosylase 2 is the highest noise and may be related

10 but has not been included.

Number of members: 189

Cyclins [1,2,3] are eukaryotic proteins which play an active role in
controlling nuclear cell division cycles. Cyclins, together with the p34

15 (cdc2) or cdk2 kinases, form the Maturation Promoting Factor (MPF). There are
two main groups of cyclins:

- G2/M cyclins, essential for the control of the cell cycle at the G2/M
(mitosis) transition. G2/M cyclins accumulate steadily during G2 and are

20 abruptly destroyed as cells exit from mitosis (at the end of the M-phase).

- G1/S cyclins, essential for the control of the cell cycle at the G1/S
(start) transition.

In most species, there are multiple forms of G1 and G2 cyclins. For example,
25 in vertebrates, there are two G2 cyclins, A and B, and at least three G1
cyclins, C, D, and E.

A cyclin homolog has also been found in herpesvirus saimiri [4].

30 The best conserved region is in the central part of the cyclins' sequences,
known as the 'cyclin-box'. From this, a 32 residue pattern has been derived.

-Consensus pattern: R-x(2)-[LIVMSA SEQ ID NO:187]-x(2)-[FYWS SEQ ID NO:40)]-[LIVM SEQ ID NO:4)]-x(8)-[LIVMFC SEQ ID NO:90)]-x(4)-[LIVMFYA SEQ ID NO:98)]-x(2)-[STAGC SEQ ID NO:45)]-[LIVMFYQ SEQ ID NO:188)]-x-[LIVMFYC SEQ ID NO:6)]-[LIVMFY SEQ ID NO:18)]-D-[RKH)]-[LIVMFYW SEQ ID NO:26)]

5

[1] Nurse P. Nature 344:503-508(1990).

[2] Norbury C., Nurse P. Curr. Biol. 1:23-24(1991).

[3] Lew D.J., Reed S.I. Trends Cell Biol. 2:77-81(1992).

[4] Nicholas J., Cameron K.R., Honess R.W. Nature 355:362-365(1992).

10

126. Cystatin domain

This is a very diverse family. Attempts to define separate subfamilies have failed. Typically, either the N-terminal or C-terminal end is very divergent. But splitting into two domains would make very short families. Cathelicidins are related to this family but have not been included. Number of members: 147

15

Inhibitors of cysteine proteases [1,2,3], which are found in the tissues and body fluids of animals, in the larva of the worm *Onchocerca volvulus* [4], as well as in plants, can be grouped into three distinct but related families:

20

- Type 1 cystatins (or stefins), molecules of about 100 amino acid residues with neither disulfide bonds nor carbohydrate groups.
- Type 2 cystatins, molecules of about 115 amino acid residues which contain one or two disulfide loops near their C-terminus.
- Kininogens, which are multifunctional plasma glycoproteins.

25

They are the precursor of the active peptide bradykinin and play a role in blood coagulation by helping to position optimally prekallikrein and factor XI next to factor XII. They are also inhibitors of cysteine proteases. Structurally, kininogens are made of three contiguous type-2 cystatin domains, followed by an additional domain (of variable length) which contains the sequence of bradykinin. The first of the three cystatin domains seems to have lost its inhibitory activity.

30

In all these inhibitors, there is a conserved region of five residues which has been proposed to be important for the binding to the cysteine proteases. The consensus pattern starts one residue before this conserved region.

-Consensus pattern: [GSTEQKRV SEQ ID NO:189]-Q-[LIVT SEQ ID NO:165)]-[VAF]-
[SAGQ SEQ ID NO:190)]-G-x-[LIVMNK SEQ ID NO:191)]-x(2)-[LIVMFY SEQ ID
NO:18)]-x-[LIVMFYA SEQ ID NO:98)]-[DENQKRHSIV SEQ ID NO:192)]

5

[1] Barrett A.J. Trends Biochem. Sci. 12:193-196(1987).

[2] Rawlings N.D., Barrett A.J. J. Mol. Evol. 30:60-71(1990).

[3] Turk V., Bode W. FEBS Lett. 285:213-219(1991).

[4] Lustigman S., Brotman B., Huima T., Prince A.M. Mol. Biochem. Parasitol. 45:65-

10 76(1991).

127. cytochrome_c (Cytochrome c)

The Pfam entry does not include all prosite members.

15 The cytochrome 556 and cytochrome c' families are
not included.

Number of members: 259

20 In proteins belonging to cytochrome c family [1], the heme group is covalently
attached by thioether bonds to two conserved cysteine residues. The consensus
sequence for this site is Cys-X-X-Cys-His and the histidine residue is one of
the two axial ligands of the heme iron. This arrangement is shared by all
proteins known to belong to cytochrome c family, which presently includes
cytochromes c, c', c1 to c6, c550 to c556, cc3/Hmc, cytochrome f and reaction
25 center cytochrome c.

-Consensus pattern: C-{CPWHF SEQ ID NO:193)}-{CPWR SEQ ID NO:194)}-C-H-
{CFYW SEQ ID NO:195)}

30 [1] Mathews F.S. Prog. Biophys. Mol. Biol. 45:1-56(1985).

128. (DAGKa) Diacylglycerol kinase accessory domain (presumed)

Diacylglycerol (DAG) is a second messenger that acts as a protein kinase C activator. This domain is assumed to be an accessory domain: its function is unknown.

[1] Sakane F, Yamada K, Kanoh H, Yokoyama C, Tanabe T, Nature 1990;344:345-348. [2] Sakane F, Imai S, Kai M, Wada I, Kanoh H, J Biol Chem 1996;271:8394-8401. [3] Schaap D, de Widt J, van der Wal J, Vandekerckhove J, van, Damme J, Gussow D, Ploegh HL, van Blitterswijk WJ, van der, Bend RL, FEBS Lett 1990;275:151-158. [4] Kanoh H, Yamada K, Sakane F, Trends Biochem Sci 1990;15:47-50.

10 129. (DAGKc) Diacylglycerol kinase catalytic domain (presumed)

Diacylglycerol (DAG) is a second messenger that acts as a protein kinase C activator. The catalytic domain is assumed from the finding of bacterial homologues.

[1] Sakane F, Yamada K, Kanoh H, Yokoyama C, Tanabe T, Nature 1990;344:345-348. [2] Sakane F, Imai S, Kai M, Wada I, Kanoh H, J Biol Chem 1996;271:8394-8401. [3] Schaap D, de Widt J, van der Wal J, Vandekerckhove J, van, Damme J, Gussow D, Ploegh HL, van Blitterswijk WJ, van der, Bend RL, FEBS Lett 1990;275:151-158. [4] Kanoh H, Yamada K, Sakane F, Trends Biochem Sci 1990;15:47-50.

20 130. D-amino acid oxidases signature(DAO)

D-amino acid oxidase (EC 1.4.3.3) (DAMOX or DAO) is an FAD flavoenzyme that catalyzes the oxidation of neutral and basic D-amino acids into their corresponding keto acids. DAOs have been characterized and sequenced in fungi and vertebrates where they are known to be located in the peroxisomes. D-aspartate oxidase (EC 1.4.3.1) (DASOX) [1] is an enzyme, structurally related to DAO, which catalyzes the same reaction but is active only toward dicarboxylic D-amino acids. In DAO, a conserved histidine has been shown [2] to be important for the enzyme's catalytic activity. The conserved region around this residue has been developed as a signature pattern for these enzymes.

30 Consensus pattern: [LIVM SEQ ID NO:4]](2)-H-[NHA]-Y-G-x-[GSA](2)-x-G-x(5)-G-x-A
[H is a probable active site residue]o-

[1] Negri A., Ceciliani F., Tedeschi G., Simonc T., Ronchi S. J. Biol. Chem. 267:11865-11871(1992).

[2] Miyano M., Fukui K., Watanabe F., Takahashi S., Tada M., Kanashiro M., Miyake Y. J. Biochem. 109:171-177(1991).

5

131. DEAD and DEAH box families ATP-dependent helicases signatures

A number of eukaryotic and prokaryotic proteins have been characterized [1,2,3] on the basis of their structural similarity. They all seem to be involved in ATP-dependent, nucleic-acid unwinding. Proteins currently known to belong to this family are: - Initiation factor eIF-4A. Found in eukaryotes, this protein is a subunit of a high molecular weight complex involved in 5'cap recognition and the binding of mRNA to ribosomes. It is an ATP-dependent RNA-helicase. - PRP5 and PRP28. These yeast proteins are involved in various ATP-requiring steps of the pre-mRNA splicing process. - P110, a mouse protein expressed specifically during spermatogenesis. - An3, a *Xenopus* putative RNA helicase, closely related to P110. - SPP81/DED1 and DBP1, two yeast proteins probably involved in pre-mRNA splicing and related to P110. - *Caenorhabditis elegans* helicase glh-1. - MSS116, a yeast protein required for mitochondrial splicing. - SPB4, a yeast protein involved in the maturation of 25S ribosomal RNA. - p68, a human nuclear antigen. p68 has ATPase and DNA-helicase activities in vitro. It is involved in cell growth and division. - Rm62 (p62), a *Drosophila* putative RNA helicase related to p68. - DBP2, a yeast protein related to p68. - DHH1, a yeast protein. - DRS1, a yeast protein involved in ribosome assembly. - MAK5, a yeast protein involved in maintenance of dsRNA killer plasmid. - ROK1, a yeast protein. - ste13, a fission yeast protein. - Vasa, a *Drosophila* protein important for oocyte formation and specification of embryonic posterior structures. - Me31B, a *Drosophila* maternally expressed protein of unknown function. - dbpA, an *Escherichia coli* putative RNA helicase. - deaD, an *Escherichia coli* putative RNA helicase which can suppress a mutation in the rpsB gene for ribosomal protein S2. - rhlB, an *Escherichia coli* putative RNA helicase. - rhIE, an *Escherichia coli* putative RNA helicase. - srmB, an *Escherichia coli* protein that shows RNA-dependent ATPase activity. It probably interacts with 23S ribosomal RNA. - *Caenorhabditis elegans* hypothetical proteins T26G10.1, ZK512.2 and ZK686.2. - Yeast hypothetical protein YHR065c. - Yeast hypothetical protein YHR169w. - Fission yeast hypothetical protein SpAC31A2.07c. - *Bacillus subtilis* hypothetical protein yxiN. All these proteins share a

number of conserved sequence motifs. Some of them are specific to this family while others are shared by other ATP-binding proteins or by proteins belonging to the helicases 'superfamily' [4,E1]. One of these motifs, called the 'D-E-A-D-box', represents a special version of the B motif of ATP-binding proteins. Some other proteins belong to a subfamily which have His instead of the second Asp and are thus said to be 'D-E-A-H-box' proteins [3,5,6,E1]. Proteins currently known to belong to this subfamily are: - PRP2, PRP16, PRP22 and PRP43. These yeast proteins are all involved in various ATP-requiring steps of the pre-mRNA splicing process. - Fission yeast prh1, which may be involved in pre-mRNA splicing. - Male-less (mle), a Drosophila protein required in males, for dosage compensation of X chromosome linked genes. - RAD3 from yeast. RAD3 is a DNA helicase involved in excision repair of DNA damaged by UV light, bulky adducts or cross-linking agents. Fission yeast rad15 (rhp3) and mammalian DNA excision repair protein XPD (ERCC-2) are the homologs of RAD3. - Yeast CHL1 (or CTF1), which is important for chromosome transmission and normal cell cycle progression in G(2)/M. - Yeast TPS1. - Yeast hypothetical protein YKL078w. - Caenorhabditis elegans hypothetical proteins C06E1.10 and K03H1.2. - Poxviruses' early transcription factor 70 Kd subunit which acts with RNA polymerase to initiate transcription from early gene promoters. - I8, a putative vaccinia virus helicase. - hrpA, an Escherichia coli putative RNA helicase. Signature patterns for both subfamilies were developed.

Consensus pattern: [LIVMF SEQ ID NO:2])(2)-D-E-A-D-[RKEN SEQ ID NO:196)]-x-[LIVMFYGSTN

Consensus pattern: [GSAH SEQ ID NO:198)]-x-[LIVMF SEQ ID NO:2])(3)-D-E-[ALIV SEQ ID NO:199)]-H-[NECR SEQ ID NO:200)]

Note: proteins belonging to this family also contain a copy of the ATP/GTP- binding motif 'A' (P-loop) (see the relevant entry <[PDOC00017](#)

[1] Schmid S.R., Linder P. Mol. Microbiol. 6:283-292(1992).

[2] Linder P., Lasko P., Ashburner M., Leroy P., Nielsen P.J., Nishi K., Schnier J., Slonimski P.P. Nature 337:121-122(1989).

[3] Wassarman D.A., Steitz J.A. Nature 349:463-464(1991).

[4] Hodgman T.C. Nature 333:22-23(1988) and Nature 333:578-578(1988) (Errata).

[5] Harosh I., Deschavanne P. Nucleic Acids Res. 19:6331-6331(1991).

[6] Koonin E.V., Senkevich T.G. J. Gen. Virol. 73:989-993(1992).

132. (DHBP_synthase) 3,4-dihydroxy-2-butanone 4-phosphate synthase

3,4-Dihydroxy-2-butanone 4-phosphate is biosynthesized from ribulose 5-phosphate and serves as the biosynthetic precursor for the xylene ring of riboflavin. Sometimes found as a bifunctional enzyme with GTP_cyclohydro2.

Richter G, Krieger C, Volk R, Kis K, Ritz H, Gotze E, Bacher A, Methods Enzymol 1997;280:374-382.

133. (DHDPS) Dihydrodipicolinate synthetase signatures

Dihydrodipicolinate synthetase (EC 4.2.1.52) (DHDPS) [1] catalyzes, in higher plants chloroplast and in many bacteria (gene dapA), the first reaction specific to the biosynthesis of lysine and of diaminopimelate. DHDPS is responsible for the condensation of aspartate semialdehyde and pyruvate by aping-pong mechanism in which pyruvate first binds to the enzyme by forming a Schiff-base with a lysine residue. Three other proteins are structurally related to DHDPS and probably also act via a similar catalytic mechanism: - Escherichia coli N-acetylneuraminate lyase (EC 4.1.3.3) (gene nanA), which catalyzes the condensation of N-acetyl-D-mannosamine and pyruvate to form N-acetylneuraminate. - Rhizobium meliloti protein mosA [3], which is involved in the biosynthesis of the rhizopine 3-o-methyl-scylo-inosamine. - Escherichia coli hypothetical protein yjhH. Two signature patterns for these enzymes were developed . The first one is centered on highly conserved region in the N-terminal part of these proteins. The second signature contains a lysine residue which has been shown, in Escherichia coli dapA [2], to be the one that forms a Schiff-base with the substrate.

Consensus pattern: [GSA]-[LIVM SEQ ID NO:4)]-[LIVMFY SEQ ID NO:18)]-x(2)-G-[ST]-[TG]-G-E-[GASNF SEQ ID NO:201)]-x(6)- [EQ] -

Consensus pattern: Y-[DNS]-[LIVMFA SEQ ID NO:81)]-P-x(2)-[ST]-x(3)-[LIVMG SEQ ID NO:202)]-x(13,14)-[LIVM SEQ ID NO:4)]- x-[SGA]-[LIVMF SEQ ID NO:2)]-K-[DEQAF SEQ ID NO:203)]-[STAC SEQ ID NO:204)] [K is involved in Schiff-base formation]-

- [1] Kaneko T., Hashimoto T., Kumpaisal R., Yamada Y. J. Biol. Chem. 265:17451-17455(1990).
- [2] Laber B., Gomis-Rueth F.-X., Romao M.J., Huber R. Biochem. J. 288:691-695(1992).
- [3] Murphy P.J., Trenz S.P., Grzemski W., de Bruijn F.J., Schell J. J. Bacteriol. 175:5193-5204 (1993).

134. (DHodehase) Dihydroorotate dehydrogenase signatures

Dihydroorotate dehydrogenase (EC 1.3.3.1) (DHodehase) catalyzes the fourth step in the de novo biosynthesis of pyrimidine, the conversion of dihydroorotate into orotate. DHodehase is a ubiquitous FAD flavoprotein. In bacteria (gene pyrD), DHodease is located on the inner side of the cytosolic membrane. In some yeasts, such as in *Saccharomyces cerevisiae* (gene URA1), it is a cytosolic protein while in other eukaryotes it is found in the mitochondria [1]. The sequence of DHodease is rather well conserved and two signature patterns were developed specific to this enzyme. The first corresponds to a region in the N-terminal section of the enzyme while the second is located in the C-terminal section and seems to be part of the FAD-binding domain.

Consensus pattern[GS]-x(4)-[GK]-[GSTA SEQ ID NO:19)]-[LIVFSTA SEQ ID NO:205)]-[GT]-x(3)-[NQR]-x-G-[NHY]-x(2)-P-[RT]

Consensus pattern[LIVM SEQ ID NO:4)](2)-[GSA]-x-G-G-[IV]-x-[STGDN SEQ ID NO:206)]-x(3)-[ACV]-x(6)-G-A

- [1] Nagy M., Lacroute F., Thomas D. Proc. Natl. Acad. Sci. U.S.A. 89:8966-8970(1992).

135. (DMRL_synthase) 6,7-dimethyl-8-ribityllumazine synthase

136. (DNA_methylase) C-5 cytosine-specific DNA methylases signatures

C-5 cytosine-specific DNA methylases (EC 2.1.1.73) (C5 Mtase) are enzymes that specifically methylate the C-5 carbon of cytosines in DNA [1,2,3]. Such enzymes are found in the proteins described below. - As a component of type II restriction-modification systems

in prokaryotes and some bacteriophages. Such enzymes recognize a specific DNA sequence where they methylate a cytosine. In doing so, they protect DNA from cleavage by type II restriction enzymes that recognize the same sequence. The sequences of a large number of type II C-5 Mtases are known. - In vertebrates, there are a number of C-5 Mtases that methylate CpG dinucleotides. The sequence of the mammalian enzyme is known. C-5 Mtases share a number of short conserved regions. Two of them were selected. The first is centered around a conserved Pro-Cys dipeptide in which the cysteine has been shown [4] to be involved in the catalytic mechanism; it appears to form a covalent intermediate with the C6 position of cytosine. The second region is located at the C-terminal extremity in type-II enzymes

Consensus pattern: [DENKS SEQ ID NO:207)]-x-[FLIV SEQ ID NO:208)]-x(2)-[GSTC SEQ ID NO:209)]-x-P-C-x(2)-[FYWLIM SEQ ID NO:210)]-S [C is the active site residue]-
Consensus pattern: [RKQGTF SEQ ID NO:211)]-x(2)-G-N-[STAG SEQ ID NO:20)]-
[LIVMF SEQ ID NO:2)]-x(3)-[LIVMT SEQ ID NO:1)]-x(3)-[LIVM SEQ ID NO:4)]- x(3)-
[LIVM SEQ ID NO:4)]-

[1] Posfai J., Bhagwat A.S., Roberts R.J. Gene 74:261-263(1988).

[2] Kumar S., Cheng X., Klimasauskas S., Mi S., Posfai J., Roberts R.J., Wilson G.G. Nucleic Acids Res. 22:1-10(1994).

[3] Lauster R., Trautner T.A., Noyer-Weidner M. J. Mol. Biol. 206:305-312(1989).

[4] Chen L., McMillan A.M., Chang W., Ezak-Nipkay K., Lane W.S., Verdine G.L. Biochemistry 30:11018-11025(1991).

137. (DNA photolyase) DNA photolyases class 2 signatures

Deoxyribodipyrimidine photolyase (EC 4.1.99.3) (DNA photolyase) [1,2] is a DNA repair enzyme. It binds to UV-damaged DNA containing pyrimidine dimers and, upon absorbing a near-UV photon (300 to 500 nm), breaks the cyclobutane ring joining the two pyrimidines of the dimer. DNA photolyase is an enzyme that requires two chromophore-cofactors for its activity: a reduced FADH₂ and either 5,10-methenyltetrahydrofolate (5,10-MTFH) or an oxidized 8-hydroxy-5-deazaflavin (8-HDF) derivative (F420). The folate or deazaflavin chromophore appears to function as an antenna, while the FADH₂ chromophore is thought to

be responsible for electron transfer. On the basis of sequence similarities[3] DNA photolyases can be grouped into two classes. The second class contains enzymes from *Myxococcus xanthus*, methanogenic archaeobacteria, insects, fish and marsupial mammals. It is not yet known what second cofactor is bound to class 2 enzymes. There are a number of conserved sequence regions in all known class 2 DNA photolyases, especially in the C-terminal part. Two of these regions were selected as signature patterns.

Consensus pattern: F-x-E-E-x-[LIVM SEQ ID NO:4)](2)-R-R-E-L-x(2)-N-F-

Consensus pattern: G-x-H-D-x(2)-W-x-E-R-x-[LIVM SEQ ID NO:4)]-F-G-K-[LIVM SEQ ID NO:4)]-R-[FY]-M-N-

[1] Sancar G.B., Sancar A. Trends Biochem. Sci. 12:259-261(1987).

[2] Jorns M.S. Biofactors 2:207-211(1990).

[3] Yasui A., Eker A.P.M., Yasuhira S., Yajima H., Kobayashi T., Takao M., Oikawa A. EMBO J. 13:6143-6151(1994).

(DNA photolyase2) DNA photolyases class 1 signatures

Deoxyribodipyrimidine photolyase (EC 4.1.99.3) (DNA photolyase) [1,2] is a DNA repair enzyme. It binds to UV-damaged DNA containing pyrimidine dimers and, upon absorbing a near-UV photon (300 to 500 nm), breaks the cyclobutane ring joining the two pyrimidines of the dimer. DNA photolyase is an enzyme that requires two chromophore-cofactors for its activity: a reduced FADH₂ and either 5,10-methenyltetrahydrofolate (5,10-MTHF) or an oxidized 8-hydroxy-5-deazaflavin (8-HDF) derivative (F420). The folate or deazaflavin chromophore appears to function as an antenna, while the FADH₂ chromophore is thought to be responsible for electron transfer. On the basis of sequence similarities[3] DNA photolyases can be grouped into two classes. The first class contains enzymes from Gram-negative and Gram-positive bacteria, the halophilic archaeobacteria *Halobacterium halobium*, fungi and plants. Class 1 enzymes bind either 5,10-MTHF (*E.coli*, fungi, etc.) or 8-HDF (*S.griseus*, *H.halobium*). This family also includes *Arabidopsis* cryptochromes 1 (CRY1) and 2 (CRY2), which are blue light photoreceptors that mediate blue light-induced gene expression. There are a number of conserved sequence regions in all known class 1 DNA photolyases, especially in the C-terminal part. Two of these regions were selected as signature patterns

Consensus pattern: T-G-x-P-[LIVM SEQ ID NO:4]](2)-D-A-x-M-[RA]-x-[LIVM SEQ ID NO:4]]-

Consensus pattern: [DN]-R-x-R-[LIVM SEQ ID NO:4]](2)-x-[STA](2)-F-[LIVMFA SEQ ID NO:81]]-x-K-x-L-x(2,3)- W-[KRQ]]-

5

[1] Sancar G.B., Sancar A. Trends Biochem. Sci. 12:259-261(1987).

[2] Jorns M.S. Biofactors 2:207-211(1990).

[3] Yasui A., Eker A.P.M., Yasuhira S., Yajima H., Kobayashi T., Takao M., Oikawa A. EMBO J. 13:6143-6151(1994).

10 [4] Lin C., Ahmad M., Cashmore A.R. Plant J. 10:893-902(1996).

138. (DNA_pol_A)

DNA polymerase family A signature

15

Replicative DNA polymerases (EC 2.7.7.7) are the key enzymes catalyzing the accurate replication of DNA. They require either a small RNA molecule or a protein as a primer for the de novo synthesis of a DNA chain. On the basis of sequence similarities a number of DNA polymerases have been grouped together [1,2,3] under the designation of DNA polymerase family A. The polymerases that belong to this family are listed below.

20

- Escherichia coli and various other bacterial polymerase I (gene polA).

- Thermus aquaticus Taq polymerase.

- Bacteriophage sp01 polymerase.

- Bacteriophage sp02 polymerase.

25

- Bacteriophage T5 polymerase.

- Bacteriophage T7 polymerase.

- Mycobacteriophage L5 polymerase.

- Yeast mitochondrial polymerase gamma (gene MIP1).

30

Five regions of similarity are found in all the above polymerases. One of these conserved regions, known as 'motif B' [1], is located in a domain which, in Escherichia coli polA, has been shown to bind deoxynucleotide triphosphate substrates; it contains a conserved tyrosine which has been shown, by photo- affinity labelling, to be in the active site; a conserved

lysine, also part of this motif, can be chemically labelled, using pyridoxal phosphate. This conserved region was used as a signature for this family of DNA polymerases.

Consensus pattern R-x(2)-[GSAV SEQ ID NO:212]-K-x(3)-[LIVMFY SEQ ID NO:18]-
 5 [AGQ]-x(2)-Y-x(2)-[GS]-x(3)- [LIVMA SEQ ID NO:30)] Sequences known to belong to this
 class detected by the pattern ALL.

[1] Delarue M., Poch O., Todro N., Moras D., Argos P. Protein Eng. 3:461-467(1990).

[2] Ito J., Braithwaite D.K. Nucleic Acids Res. 19:4045-4057(1991).

10 [3] Braithwaite D.K., Ito J. Nucleic Acids Res. 21:787-802(1993).

139. DNA_pol_viral_C

DNA polymerase (viral) C-terminal domain

15 Number of members: 128

140. (DNA_topoisII)

DNA topoisomerase II signature

20 DNA topoisomerase I (EC 5.99.1.2) [1,2,3,4,E1] is one of the two types of enzyme that
 catalyze the interconversion of topological DNA isomers. Type II topoisomerases are ATP-
 dependent and act by passing a DNA segment through a transient double-strand break.
 Topoisomerase II is found in phages, archaeobacteria, prokaryotes, eukaryotes, and in African
 Swine Fever virus (ASF). In bacteriophage T4 topoisomerase II consists of three subunits
 25 (the product of genes 39, 52 and 60). In prokaryotes and in archaeobacteria the enzyme,
 known as DNA gyrase, consists of two subunits (genes gyrA and gyrB [E2]). In some
 bacteria, a second type II topoisomerase has been identified; it is known as topoisomerase IV
 and is required for chromosome segregation, it also consists of two subunits (genes parC and
 parE). In eukaryotes, type II topoisomerase is a homodimer.

30

There are many regions of sequence homology between the different subtypes of
 topoisomerase II. The relation between the different subunits is shown in the following
 representation:

<-----About-1400-residues----->

[-----Protein 39-*-----][----Protein 52----] Phage T4

5 [-----gyrB-----*-----][-----gyrA-----] Prokaryote II

Archaeobacteria

[-----parE-----*-----][-----parD-----] Prokaryote IV

[-----*-----] Eukaryote and

ASF

10 '*': Position of the pattern.

As a signature pattern for this family of proteins, a region that contains a highly conserved pentapeptide was selected. The pattern is located in gyrB, in parE, and in protein 39 of phage T4 topoisomerase.

15

Consensus pattern[LIVMA SEQ ID NO:30])-x-E-G-[DN]-S-A-x-[STAG SEQ ID NO:20)]

Sequences known to belong to this class detected by the pattern ALL.

[1] Sternglanz R. Curr. Opin. Cell Biol. 1:533-535(1990).

20 [2] Bjornsti M.-A. Curr. Opin. Struct. Biol. 1:99-103(1991).

[3] Sharma A., Mondragon A. Curr. Opin. Struct. Biol. 5:39-47(1995).

[4] Roca J. Trends Biochem. Sci. 20:156-160(1995).

25 141. (DSPc) Tyrosine specific protein phosphatases signature and profiles

Tyrosine specific protein phosphatases (EC 3.1.3.48) (PTPase) [1 to 5] are enzymes that catalyze the removal of a phosphate group attached to a tyrosine residue. These enzymes are very important in the control of cell growth, proliferation, differentiation and transformation. Multiple forms of PTPase have been characterized and can be classified into two categories:

30 soluble PTPases and transmembrane receptor proteins that contain PTPase domain(s). The currently known PTPases are listed below: Soluble PTPases. - PTPN1 (PTP-1B). - PTPN2 (T-cell PTPase; TC-PTP). - PTPN3 (H1) and PTPN4 (MEG), enzymes that contain an N-terminal band 4.1- like domain (see <PDOC00566>) and could act at junctions between the

membrane and cytoskeleton. - PTPN5 (STEP). - PTPN6 (PTP-1C; HCP; SHP) and PTPN11 (PTP-2C; SH-PTP3; Syp), enzymes which contain two copies of the SH2 domain at its N-terminal extremity. The *Drosophila* protein corkscrew (gene *csw*) also belongs to this subgroup. - PTPN7 (LC-PTP; Hematopoietic protein-tyrosine phosphatase; HePTP). - PTPN8 (70Z-PEP). - PTPN9 (MEG2). - PTPN12 (PTP-G1; PTP-P19). - Yeast PTP1. - Yeast PTP2 which may be involved in the ubiquitin-mediated protein degradation pathway. - Fission yeast *pyp1* and *pyp2* which play a role in inhibiting the onset of mitosis. - Fission yeast *pyp3* which contributes to the dephosphorylation of *cdc2*. - Yeast CDC14 which may be involved in chromosome segregation. - *Yersinia* virulence plasmid PTPases (gene *yopH*). - *Autographa californica* nuclear polyhedrosis virus 19 Kd PTPase. Dual specificity PTPases. - DUSP1 (PTPN10; MAP kinase phosphatase-1; MKP-1); which dephosphorylates MAP kinase on both Thr-183 and Tyr-185. - DUSP2 (PAC-1), a nuclear enzyme that dephosphorylates MAP kinases ERK1 and ERK2 on both Thr and Tyr residues. - DUSP3 (VHR). - DUSP4 (HVH2). - DUSP5 (HVH3). - DUSP6 (Pyst1; MKP-3). - DUSP7 (Pyst2; MKP-X). - Yeast MSG5, a PTPase that dephosphorylates MAP kinase FUS3. - Yeast YVH1. - *Vaccinia* virus H1 PTPase; a dual specificity phosphatase. Receptor PTPases. Structurally, all known receptor PTPases, are made up of a variable length extracellular domain, followed by a transmembrane region and a C-terminal catalytic cytoplasmic domain. Some of the receptor PTPases contain fibronectin type III (FN-III) repeats, immunoglobulin-like domains, MAM domains or carbonic anhydrase-like domains in their extracellular region. The cytoplasmic region generally contains two copies of the PTPase domain. The first seems to have enzymatic activity, while the second is inactive but seems to affect substrate specificity of the first. In these domains, the catalytic cysteine is generally conserved but some other, presumably important, residues are not. In the following table, the domain structure of known receptor PTPases is shown:

receptor PTPases	Extracellular	Intracellular
Ig FN-3	0 2 0 0 2	0 2 0 0 2
Leukocyte common antigen (LCA) (CD45)	0 2 0 0 2	0 2 0 0 2
Leukocyte antigen related (LAR)	3 8 0 0 2	0 2 0 0 2
<i>Drosophila</i> DLAR	3 9 0 0 2	0 2 0 0 2
<i>Drosophila</i> DPTP	2 2 0 0 2	0 2 0 0 2
PTP-alpha (LRP)	0 0 0 0 2	0 16 0 0 1
PTP-beta	0 16 0 0 1	0 1 1 0 2
PTP-gamma	0 1 1 0 2	0 7 0 0 2
PTP-delta	0 7 0 0 2	0 0 0 0 2
PTP-epsilon	0 0 0 0 2	1 4 0 1 2
PTP-kappa	1 4 0 1 2	1 4 0 1 2
PTP-mu	1 4 0 1 2	0 1 1 0 2
PTP-zeta	0 1 1 0 2	0 1 1 0 2

PTPase domains consist of about 300 amino acids. There are two conserved cysteines, the second one has been shown to be absolutely required for activity. Furthermore, a number of conserved residues in its immediate vicinity have also been shown to be important. A signature pattern for PTPase domains was derived centered on the active site cysteine. There are three profiles for

PTPases, the first one spans the complete domain and is not specific to any subtype. The second profile is specific to dual-specificity PTPases and the third one to the PTP subfamily

Consensus pattern: [LIVMF SEQ ID NO:2)]-H-C-x(2)-G-x(3)-[STC]-[STAGP SEQ ID
5 NO:213)]-x-[LIVMFY SEQ ID NO:18)] [C is the active site residue]-

[1] Fischer E.H., Charbonneau H., Tonks N.K. Science 253:401-406(1991).

[2] Charbonneau H., Tonks N.K. Annu. Rev. Cell Biol. 8:463-493(1992).

[3] Trowbridge I.S. J. Biol. Chem. 266:23517-23520(1991).

10 [4] Tonks N.K., Charbonneau H. Trends Biochem. Sci. 14:497-500(1989).

[5] Hunter T. Cell 58:1013-1016(1989).

142. (DUF10) Uncharacterized protein family UPF0076 signature

15 The following uncharacterized proteins have been shown [1] to share regions of similarities: -
Goat antigen UK114, a human homolog and the rat corresponding protein which is known as
perchloric acid soluble protein (PSP1). PSP1 [2] may inhibit an initiation stage of cell-free
protein synthesis. - Mouse heat-responsive protein HRSP12. - Yeast chromosome V
hypothetical protein YER057c. - Yeast chromosome IX hypothetical protein YIL051c. -
20 Caenorhabditis elegans hypothetical protein C23G10.2. - Escherichia coli hypothetical
protein ycdK. - Escherichia coli hypothetical protein yhaR. - Escherichia coli hypothetical
protein yjgF and HI0719, the corresponding Haemophilus influenzae protein. - Escherichia
coli hypothetical protein yoaB. - Bacillus subtilis hypothetical protein yabJ. - Haemophilus
influenzae hypothetical protein HI1627. - Helicobacter pylori hypothetical protein HP0944. -
25 Lactococcus lactis aldR. - Myxococcus xanthus dfrA. - Synechocystis strain PCC 6803
hypothetical protein slr0709. - Rhizobium strain NGR234 symbiotic plasmid hypothetical
protein y4sK. - Pyrococcus horikoshii hypothetical protein PH0854. These are small proteins
of around 15 Kd whose sequence is highly conserved. As a signature pattern, a well conserved
region located in the C-terminal part of these proteins was selected.

30 Consensus pattern: [PA]-[ASTPV SEQ ID NO:214)]-R-[SACVF SEQ ID NO:215)]-x-
[LIVMFY SEQ ID NO:18)]-x(2)-[GSAKR SEQ ID NO:216)]-x-[LMVA SEQ ID NO:217)]-
x(5,8)-[LIVM SEQ ID NO:4)]-E-[MI]-

[1] Bairoch A. Unpublished observations (1995).

[2] Oka T., Tsuji H., Noda C., Sakai K., Hong Y.-M., Suzuki I., Munoz S., Natori Y. J. Biol. Chem. 270:30060-30067(1995).

5

143. (DUF3)Domain of Unknown Function 3

Domain apparently occurring exclusively in eubacteria. Unknown function.

10

144. (DUF6) Integral membrane protein

This family includes many hypothetical membrane proteins of unknown function. Many of the proteins contain two copies of the aligned region.

15

145. (DUF7) Integral membrane protein

This family includes many hypothetical membrane proteins of unknown function. Swiss:P14502 has been implicated in resistance to ethidium bromide.

20

146. (DapB) Dihydrodipicolinate reductase signature

Dihydrodipicolinate reductase (EC 1.3.1.26) catalyzes the second step in the biosynthesis of diaminopimelic acid and lysine, the NAD or NADP-dependent reduction of 2,3-dihydrodipicolinate into 2,3,4,5-tetrahydrodipicolinate. This enzyme is present in bacteria (gene dapB) and higher plants. As a signature pattern the best conserved region in this enzyme was selected. It is located in the central section and is part of the substrate-binding region [1].

25

30 Consensus pattern: E-[IV]-x-E-x-H-x(3)-K-x-D-x-P-S-G-T-A-

[1] Scapin G., Blanchard J.S., Sacchettini J.C. Biochemistry 34:3502-3512(1995).

147. DedA family

This family combines the DedA related proteins and YIAN/YGIK family. Members of this family are not functionally characterised. These proteins contain multiple predicted transmembrane regions.

148. DegT/DnrJ/EryC1/StrS family

The members of this family exhibit some characteristics of the sensor protein of two-component signal transduction systems, however none of the members show any sequence similarity to these protein kinases. The members of this family do have the typical helix-turn-helix motif of DNA binding proteins.

[1] Stutzman-Engwall KJ, Otten SL, Hutchinson CR, J Bacteriol 1992;174:144-154.

149. (Desaturase) Fatty acid desaturases signatures

Fatty acid desaturases (EC 1.14.99.-) are enzymes that catalyze the insertion of a double bond at the delta position of fatty acids. There seems to be two distinct families of fatty acid desaturases which do not seem to be evolutionary related. Family 1 is composed of: -

Stearoyl-CoA desaturase (SCD) (EC 1.14.99.5) [1]. SCD is a key regulatory enzyme of unsaturated fatty acid biosynthesis. SCD introduces a cis double bond at the delta(9) position of fatty acyl-CoA's such as palmitoleoyl- and oleoyl-CoA. SCD is a membrane-bound enzyme that is thought to function as a part of a multienzyme complex in the endoplasmic reticulum of vertebrates and fungi. As a signature pattern for this family a conserved region in the C-terminal part of these enzymes was selected, this region is rich in histidine residues and in aromatic residues. Family 2 is composed of: - Plants stearoyl-acyl-carrier-protein desaturase (EC 1.14.99.6) [2], these enzymes catalyze the introduction of a double bond at the delta(9) position of stearoyl-ACP to produce oleoyl-ACP. This enzyme is responsible for the conversion of saturated fatty acids to unsaturated fatty acids in the synthesis of vegetable oils. - Cyanobacteria desA [3] an enzyme that can introduce a second cis double bond at the delta(12) position of fatty acid bound to membranes glycerolipids. DesA is involved in chilling tolerance; the phase transition temperature of lipids of cellular membranes being dependent on the degree of unsaturation of fatty acids of the membrane lipids. As a signature

pattern for this family a conserved region in the C-terminal part of these enzymes was selected.

Consensus pattern: G-E-x-[FY]-H-N-[FY]-H-H-x-F-P-x-D-Y-

5 Consensus pattern: [ST]-[SA]-x(3)-[QR]-[LI]-x(5,6)-D-Y-x(2)-[LIVMFYW SEQ ID NO:26)]-[LIVM SEQ ID NO:4)]- [DE]-

[1] Kaestner K.H., Ntambi J.M., Kelly T.J. Jr., Lane M.D. J. Biol. Chem. 264:14755-14761(1989).

10 [2] Shanklin J., Somerville C.R. Proc. Natl. Acad. Sci. U.S.A. 88:2510-2514(1991).

[3] Wada H., Gombos Z., Murata N. Nature 347:200-203(1990).

150. Dihydroorotase signatures

15 Dihydroorotase (EC 3.5.2.3) (DHOase) catalyzes the third step in the de novo biosynthesis of pyrimidine, the conversion of ureidosuccinic acid (N-carbamoyl-L-aspartate) into dihydroorotate. Dihydroorotase binds a zinc ion which is required for its catalytic activity [1]. In bacteria, DHOase is a dimer of identical chains of about 400 amino-acid residues (gene pyrC). In higher eukaryotes, DHOase is part of a large multi-functional protein known as

20 'rudimentary' in Drosophila and CAD in mammals and which catalyzes the first three steps of pyrimidine biosynthesis [2]. The DHOase domain is located in the central part of this polyprotein. In yeasts, DHOase is encoded by a monofunctional protein (gene URA4). However, a defective DHOase domain [3] is found in a multifunctional protein (gene URA2) that catalyzes the first two steps of pyrimidine biosynthesis. The comparison of

25 DHOase sequences from various sources shows [4] that there are two highly conserved regions. The first located in the N-terminal extremity contains two histidine residues suggested [3] to be involved in binding the zinc ion. The second is found in the C-terminal part. Signature patterns for both regions have been developed. Allantoinase (EC 3.5.2.5) is the enzyme that hydrolyzes allantoin into allantoate. In yeast (gene DAL1) [5], it is the first

30 enzyme in the allantoin degradation pathway; in amphibians [6] and fish it catalyzes the second step in the degradation of uric acid. The sequence of allantoinase is evolutionary related to that of DHOases.

Consensus pattern: D-[LIVMFYWSAP SEQ ID NO:218)]-H-[LIVA SEQ ID NO:219)]-H-[LIVF SEQ ID NO:127)]-[RN]-x-[PGANF SEQ ID NO:220)] [The two H's are probable zinc ligands]-

Consensus pattern: [GA]-[ST]-D-x-A-P-H-x(4)-K-

5

[1] Brown D.C., Collins K.D. J. Biol. Chem. 266:1597-1604(1991).

[2] Davidson J.N., Chen K.C., Jamison R.S., Musmanno L.A., Kern C.B. BioEssays 15:157-164(1993).

[3] Souciet J.-L., Nagy M., Le Gouar M., Lacroute F., Potier S. Gene 79:59-70(1989).

10 [4] Guyonvarch A., Nguyen-Juilleret M., Hubert J.-C., Lacroute F. Mol. Gen. Genet. 212:134-141(1988).

[5] Buckholz R.G., Cooper T.G. Yeast 7:913-923(1991).

[6] Hayashi S., Jain S., Chu R., Alvares K., Xu B., Erfurth F., Usuda N., Rao M.S., Reddy S.K., Noguchi T., Reddy J.K., Yeldandi A.Y. J. Biol. Chem. 269:12269-12276(1994).

15

151. dnaJ domains signatures and profile

The prokaryotic heat shock protein dnaJ interacts with the chaperone hsp70-like dnaK protein [1]. Structurally, the dnaJ protein consists of an N- terminal conserved domain (called 'J' domain) of about 70 amino acids, a glycine-rich region ('G' domain') of about 30 residues, a central domain containing four repeats of a CXXCXGXG motif ('CRR' domain) and a C-terminal region of 120 to 170 residues. Such a structure is shown in the following schematic representation:

25 +-----+-----+-----+-----+-----+-----+-----+-----+ | N-terminal | |
Gly-R | | CXXCXGXG | C-terminal | +-----+-----+-----+-----+-----+-----+
-----+

It has been shown [2] that the 'J' domain as well as the 'CRR' domain are also found in other prokaryotic and eukaryotic proteins which are listed below.

a) Proteins containing both a 'J' and a 'CRR' domain:

30

- Yeast protein MAS5/YDJ1 which seems to be involved in mitochondrial protein import.
- Yeast protein MDJ1, involved in mitochondrial biogenesis and protein folding.
- Yeast protein SCJ1, involved in protein sorting.

- Yeast protein XDJ1.
- Plants dnaJ homologs (from leek and cucumber).
- Human HDJ2, a dnaJ homolog of unknown function.
- Yeast hypothetical protein YNL077w.

5 b) Proteins containing a 'J' domain without a 'CRR' domain:

- Rhizobium fredii nolC, a protein involved in cultivar-specific nodulation of soybean.
- Escherichia coli cbpA [3], a protein that binds curved DNA.
- Yeast protein SEC63/NPL1, important for protein assembly into the endoplasmic
10 reticulum and the nucleus.
- Yeast protein SIS1, required for nuclear migration during mitosis.
- Yeast protein CAJ1.
- Yeast hypothetical protein YFR041c.
- Yeast hypothetical protein YIR004w.
- Yeast hypothetical protein YJL162c.
15
- Plasmodium falciparum ring-infected erythrocyte surface antigen (RESA). RESA, whose function is not known, is associated with the membrane skeleton of newly invaded erythrocytes.
- Human HDJ1.
- Human HSJ1, a neuronal protein.
20
- Drosophila cysteine-string protein (csp).

A signature pattern for the 'J' domain was developed, based on conserved positions in the C-terminal half of this domain. A pattern for the 'CRR' domain, based on the first two copies of that motif was also developed. A profile for the 'J' domain was also developed.

25 Consensus pattern: [FY]-x(2)-[LIVMA SEQ ID NO:30)]-x(3)-[FYWHNT SEQ ID NO:221)]-[DENQSA SEQ ID NO:222)]-x-L-x-[DN]-x(3)- [KR]-x(2)-[FYI]-

Consensus pattern: C-[DEGSTHKR SEQ ID NO:223)]-x-C-x-G-x-[GK]-[AGSDM SEQ ID NO:224)]-x(2)-[GSNKR SEQ ID NO:225)]-x(4,6)-C- x(2,3)-C-x-G-x-G-

30 [1] Cyr D.M., Langer T., Douglas M.G. Trends Biochem. Sci. 19:176-181(1994).

[2] Bork P., Sander C., Valencia A., Bukau B. Trends Biochem. Sci. 17:129-129(1992).

[3] Ueguchi C., Kaneda M., Yamada H., Mizuno T. Proc. Natl. Acad. Sci. U.S.A. 91:1054-1058(1994).

5 152.

153. Dwarfins

This family known as the dwarfins also includes the drosophila protein MAD. The N-terminus of MAD can bind to DNA [2].

10 [1] Yingling JM, Das P, Savage C, Zhang M, Padgett RW, Wang XF, Proc Natl Acad Sci U S A 1996;93:8940-8944. [2] Kim J, Johnson K, Chen HJ, Carroll S, Laughon A, Nature 1997;388:304-308.

15 154. Dynein light chain type 1 signature

Dynein is a multisubunit microtubule-dependent motor enzyme that acts as the force generating protein of eukaryotic cilia and flagella. The cytoplasmic isoform of dynein acts as a motor for the intracellular retrograde motility of vesicles and organelles along microtubules. Dynein is composed of a number of ATP-binding large subunits, intermediate size subunits and small subunits. Among the small subunits, there is a family [1,2] of highly conserved proteins which consist of: - Chlamydomonas reinhardtii flagellar outer arm dynein 8 Kd and 11 Kd light chains. - Higher eukaryotes cytoplasmic dynein light chain 1. - Yeast cytoplasmic dynein light chain 1 (gene DYN2 or SLC1). - Caenorhabditis elegans hypothetical dynein light chains M18.2 and T26A5.9. These proteins are have from 89 to 120 amino acids. As a signature pattern, A highly conserved region was selected.

20
25

Consensus pattern: H-x-I-x-G-[KR]-x-F-[GA]-S-x-V-[ST]-[HY]-E -

[1] King S.M., Patel-King R.S. J. Biol. Chem. 270:11445-11452(1995).

30 [2] Dick T., Ray K., Salz H.K., Chia W. Mol. Cell. Biol. 16:1966-1977(1996).

155. dUTPase

dUTPase hydrolyzes dUTP to dUMP and pyrophosphate.

[1] Cedergren-Zeppezauer ES, Larsson G, Nyman PO, Dauter Z, Wilson KS, Nature 1992;355:740-743. [2] Mol CD, Harris JM, McIntosh EM, Tainer JA, Structure 1996;4:1077-1092.

5

156. (dCMP cyt deam) Cytidine and deoxycytidylate deaminases zinc-binding region signature

Cytidine deaminase (EC 3.5.4.5) (cytidine aminohydrolase) catalyzes the hydrolysis of
 10 cytidine into uridine and ammonia while deoxycytidylatedeaminase (EC 3.5.4.12) (dCMP
 deaminase) hydrolyzes dCMP into dUMP. Both enzymes are known to bind zinc and to
 require it for their catalytic activity[1,2]. These two enzymes do not share any sequence
 similarity with the exception of a region that contains three conserved histidine and cysteine
 residues which are thought to be involved in the binding of the catalytic zincion. Such a
 15 region is also found in other proteins [3,4]: - Yeast cytosine deaminase (EC 3.5.4.1) (gene
 FCY1) which transforms cytosine into uracil. - Mammalian apolipoprotein B mRNA editing
 protein, responsible for the postranscriptional editing of a CAA codon into a UAA (stop)
 codon in the APOB mRNA. - Riboflavin biosynthesis protein ribG, which converts 2,5-
 diamino-6- (ribosylamino)-4(3H)-pyrimidinone 5'-phosphate into 5-amino-6- (ribosylamino)-
 20 2,4(1H,3H)-pyrimidinedione 5'-phosphate. - Bacillus cereus blasticidin-S deaminase (EC
3.5.4.23), which catalyzes the deamination of the cytosine moiety of the antibiotics
 blasticidin S, cytomycin and acetylblasticidin S. - Bacillus subtilis protein comEB. This
 protein is required for the binding and uptake of transforming DNA. - Bacillus subtilis
 hypothetical protein yaaJ. - Escherichia coli hypothetical protein yfhC. - Yeast hypothetical
 25 protein YJL035c. A signature pattern for this zinc-binding region was derived.

Consensus pattern: [CH]-[AGV]-E-x(2)-[LIVMFGAT SEQ ID NO:226)]-[LIVM SEQ ID
 NO:4)]-x(17,33)-P-C-x(2,8)-C- x(3)-[LIVM SEQ ID NO:4)] [The C's and H are zinc ligands

30 [1] Yang C., Carlow D., Wolfenden R., Short S.A. Biochemistry 31:4168-4174(1992).
 [2] Moore J.T., Silversmith R.E., Maley G.F., Maley F. J. Biol. Chem. 268:2288-
 2291(1993).
 [3] Reizer J., Buskirk S., Bairoch A., Reizer A., Saier M.H. Jr. Protein Sci. 3:853-856(1994).

[4] Bhattacharya S., Navaratnam N., Morrison J.R., Scott J., Taylow W.R. Trends Biochem. Sci. 19:105-106(1994).

5 157. Dehydrins signatures

A number of proteins are produced by plants that experience water-stress. Water-stress takes place when the water available to a plant falls below a critical level. The plant hormone abscisic acid (ABA) appears to modulate the response of plant to water-stress. Proteins that are expressed during water-stress are called dehydrins [1,2] or LEA group 2 proteins [3]. The proteins that belong to this family are listed below.

- Arabidopsis thaliana XERO 1, XERO 2 (LTI30), RAB18, ERD10 (LTI45) ERD14 and COR47.
- Barley dehydrins B8, B9, B17, and B18.
- Cotton LEA protein D-11.
- Craterostigma plantagineum dessication-related proteins A and B.
- Maize dehydrin M3 (RAB-17).
- Pea dehydrins DHN1, DHN2, and DHN3.
- Radish LEA protein.
- Rice proteins RAB 16B, 16C, 16D, RAB21, and RAB25.
- Tomato TAS14.
- Wheat dehydrin RAB 15 and cold-shock protein cor410, cs66 and cs120.

Dehydrins share a number of structural features. One of the most notable features is the presence, in their central region, of a continuous run of five to nine serines followed by a cluster of charged residues. Such a region has been found in all known dehydrins so far with the exception of pea dehydrins. A second conserved feature is the presence of two copies of alysine-rich octapeptide; the first copy is located just after the cluster of charged residues that follows the poly-serine region and the second copy is found at the C-terminal extremity. Signature patterns for both regions were derived.

Consensus pattern: S(5)-[DE]-x-[DE]-G-x(1,2)-G-x(0,1)-[KR](4

Consensus pattern: [KR]-[LIM]-K-[DE]-K-[LIM]-P-G-

[1] Close T.J., Kortt A.A., Chandler P.M. Plant Mol. Biol. 13:95-108(1989).

[2] Robertson M., Chandler P.M. Plant Mol. Biol. 19:1031-1044(1992).

[3] Dure L. III, Crouch M., Harada J., Ho T.-H. D., Mundy J., Quatrano R., Thomas T., Sung Z.R. *Plant Mol. Biol.* 12:475-486(1989).

5 158. (deoR) Bacterial regulatory proteins, deoR family signature

The many bacterial transcription regulation proteins which bind DNA through a helix-turn-helix' motif can be classified into subfamilies on the basis of sequence similarities. One of these subfamilies groups the following proteins[1,2]: - accR, the *Agrobacterium tumefaciens* plasmid pTiC58 repressor of opine catabolism and conjugal transfer. - agaR, the *Escherichia coli* aga operon putative repressor. - deoR, the *Escherichia coli* deoxyribose operon repressor. - fucR, the *Escherichia coli* L-fucose operon activator. - gatR, the *Escherichia coli* galactitol operon repressor. - glpR, the *Escherichia coli* glycerol-3-phosphate regulon repressor. - gutR (or srlR), the *Escherichia coli* glucitol operon repressor. - iolR, from *Bacillus subtilis*. - lacR, the *streptococci* lactose phosphotransferase system repressor. - spoIIID, the *Bacillus subtilis* transcription regulator of the sigK gene. - yfjR, an *Escherichia coli* hypothetical protein. - ygbI, an *Escherichia coli* hypothetical protein. - yihW, an *Escherichia coli* hypothetical protein. - yjfQ, an *Escherichia coli* hypothetical protein. - yjhJ, an *Escherichia coli* hypothetical protein. The 'helix-turn-helix' DNA-binding motif of these proteins is located in the N-terminal part of the sequence. The pattern used to detect these proteins starts fourteen residues before the HTH motif and ends one residue after it.

Consensus pattern: R-x(3)-[LIVM SEQ ID NO:4)]-x(3)-[LIVM SEQ ID NO:4)]-x(16,17)-[STA]-x(2)-T-[LIVMA SEQ ID NO:30)]- [RH]-[KRNA SEQ ID NO:227)]-D-[LIVMF SEQ ID NO:2)]-

[1] von Bodman S., Hayman G.T., Farrand S.K. *Proc. Natl. Acad. Sci. U.S.A.* 89:643-647(1992).

[2] Bairoch A. Unpublished observations (1993).

159. dsrm

Double-stranded RNA binding motif

[1] Burd CG, Dreyfuss G; Medline: 94310455, Conserved structures and diversity of functions of RNA-binding proteins. Science 1994;265:615-621.

Sequences gathered for seed by HMM_iterative_training Putative motif shared by proteins that bind to dsRNA. At least some DSRM proteins seem to bind to specific RNA targets. Exemplified by Staufen, which is involved in localization of at least five different mRNAs in the early Drosophila embryo. Also by interferon-induced protein kinase in humans, which is part of the cellular response to dsRNA.

Number of members: 116

160. Dynamin family signature

Dynamin [1,2] is a microtubule-associated force-producing protein of 100 Kd which is involved in the production of microtubule bundles and which is able to bind and hydrolyze GTP. Dynamin is structurally related to the following proteins: - Drosophila shibire protein (gene shi) [3]. Shibire is, very probably, the Drosophila cognate of mammalian dynamin. It seems to provide the motor for vesicular transport during endocytosis. - Yeast vacuolar sorting protein VPS1 (or SPO15) [4], a protein which could also be involved in microtubule-associated motility. - Yeast protein MGM1 [5], which is required for mitochondrial genome maintenance. - Yeast protein DNM1, which is involved in endocytosis. - Interferon induced Mx proteins [6,7]. Interferon alpha or beta induce the synthesis of a family of closely related proteins. Most of these proteins are known to confer resistance to influenza viruses and/or rhabdoviruses on transfected mammalian cell in culture. The three motifs found in all GTP-binding proteins are located in the N-terminal part of these proteins. The signature pattern that was developed for these proteins is based on a highly conserved region downstream of the ATP/GTP-binding motif 'A' (P-loop) (see <PDOC00017>).-

Consensus pattern: L-P-[RK]-G-[STN]-[GN]-[LIVM SEQ ID NO:4]-V-T-R-

[1] Vallee R.B., Shpetner H.S. Annu. Rev. Biochem. 59:909-932(1990).

[2] Obar R.A., Collins C.A., Hammarback J.A., Shpetner H.S., Vallee R.B. Nature 347:256-261(1990).

[3] van der Blik A., Meyerowitz E.M. Nature 351:411-414(1991).

[4] Rothman J.H., Raymond C.K., Gilbert T., O'Hara P.J., Stevens T.H. Cell 61:1063-1074(1990).

[5] Jones B.A., Fangman W.L. Genes Dev. 6:380-389(1992).

5 [6] Arnheiter H., Meier E. New Biol. 2:851-857(1990).

[7] Staeheli P., Pitossi F., Pavlovic J. Trends Cell Biol. 3:268-272(1993).

161. (dynamin_2) Dynamin central region

10 This region lies between the GTPase domain, see dynamin, and the pleckstrin homology (PH) domain.

162. E1-E2 ATPases phosphorylation site

15 E1-E2 ATPases (also known as P-type) are cation transport ATPases which form an aspartyl phosphate intermediate in the course of ATP hydrolysis. ATPases which belong to this family are listed below [1,2,3]. - Fungal and plant plasma membrane (H⁺) ATPases [reviewed in 4]. - Vertebrate (Na⁺, K⁺) ATPases (sodium pump) [reviewed in 5,6]. - Gastric (K⁺, H⁺) ATPases (proton pump). - Calcium (Ca⁺⁺) ATPases (calcium pump) from the sarcoplasmic
20 reticulum (SR), the endoplasmic reticulum (ER) and the plasma membrane. - Copper (Cu⁺⁺) ATPases (copper pump) which are involved in two human genetic disorders: Menkes syndrome and Wilson disease [7]. - Bacterial potassium (K⁺) ATPases. - Bacterial cadmium efflux (Cd⁺⁺) ATPases [reviewed in 8]. - Bacterial magnesium (Mg⁺⁺) ATPases. - A
25 probable cation ATPase from Leishmania. - fixI, a probable cation ATPase from Rhizobium meliloti, involved in nitrogen fixation. The region around the phosphorylated aspartate residue is perfectly conserved in all these ATPases and can be used as a signature pattern.

Consensus pattern: D-K-T-G-T-[LI]-[TI] [D is phosphorylated]

30 [1] Green N.M., McLennan D.H. Biochem. Soc. Trans. 17:819-822(1989).

[2] Green N.M. Biochem. Soc. Trans. 17:970-972(1989).

[3] Fagan M.J., Saier M.H. Jr. J. Mol. Evol. 38:57-99(1994).

[4] Serrano R. Biochim. Biophys. Acta 947:1-28(1988).

- [5] Fambrough D.M. Trends Neurosci. 11:325-328(1988).
[6] Sweadner K.J. Biochim. Biophys. Acta 988:185-220(1989).
[7] Bull P.C., Cox D.W. Trends Genet. 10:246-251(1994).
[8] Silver S., Nucifora G., Chu L., Misra T.K. Trends Biochem. Sci. 14:76-80(1989).

5

163. E1_N

E1 Protein, N terminal domain

Number of members: 90

10

164. (E1_dehydrog) Dehydrogenase E1 component

This family uses thiamine pyrophosphate as a cofactor. This family includes pyruvate dehydrogenase, 2-oxoglutarate dehydrogenase and 2-oxoisovalerate dehydrogenase.

15

165. (ECH) Enoyl-CoA hydratase/isomerase signature

Enoyl-CoA hydratase (EC 4.2.1.17) (ECH) [1] and 3-2trans-enoyl-CoA isomerase(EC 5.3.3.8) (ECI) [2] are two enzymes involved in fatty acid metabolism. ECH catalyzes the hydration of 2-trans-enoyl-CoA into 3-hydroxyacyl-CoA and ECI shifts the 3- double bond of the intermediates of unsaturated fatty acid oxidation to the 2-trans position. Most eukaryotic cells have two fatty-acid beta-oxidation systems, one located in mitochondria and the other in peroxisomes. In mitochondria, ECH and ECI are separate yet structurally related monofunctional enzymes. Peroxisomes contain a trifunctional enzyme [3] consisting of an N-terminal domain that bears both ECH and ECI activity, and a C-terminal domain responsible for 3-hydroxyacyl-CoA dehydrogenase (HCDH) activity. In Escherichia coli (gene fadB) and Pseudomonas fragi (gene faoA), ECH and ECI are also part of a multifunctional enzyme which contains both a HCDH and a3-hydroxybutyryl-CoA epimerase domain [4].A number of other proteins have been found to be evolutionary related to the ECH/ECI enzymes or domains: - 3-hydroxybutyryl-coa dehydratase (EC 4.2.1.55) (crotonase), a bacterial enzyme involved in the butyrate/butanol-producing pathway. - Naphthoate synthase (EC 4.1.3.36) (DHNA synthetase) (gene menB) [5], a bacterial enzyme involved in the biosynthesis of menaquinone (vitamin K2). DHNA synthetase converts O-succinyl-benzoyl-CoA (OSB-

20

25

30

CoA) to 1,4-dihydroxy- 2-naphthoic acid (DHNA). - 4-chlorobenzoate dehalogenase (EC 3.8.1.6) [6], a *Pseudomonas* enzyme which catalyzes the conversion of 4-chlorobenzoate-CoA to 4-hydroxybenzoate-CoA. - A *Rhodobacter capsulatus* protein of unknown function (ORF257) [7]. - *Bacillus subtilis* putative polyketide biosynthesis proteins pksH and pksI. -
 5 *Escherichia coli* carnitine racemase (gene caiD) [8]. - *Escherichia coli* hypothetical protein ygfG. - Yeast hypothetical protein YDR036c. As a signature pattern for these enzymes, a conserved region rich in glycine and hydrophobic residues was selected.

Consensus pattern: [LIVM SEQ ID NO:4)]-[STA]-x-[LIVM SEQ ID NO:4)]-[DENQRHSTA
 10 SEQ ID NO:228)]-G-x(3)-[AG](3)-x(4)- [LIVMST SEQ ID NO:48)]-x-[CSTA SEQ ID NO:229)]-[DQHP SEQ ID NO:230)]-[LIVMFY SEQ ID NO:18)]-

[1] Minami-Ishii N., Taketani S., Osumi T., Hashimoto T. Eur. J. Biochem. 185:73-78(1989).

15 [2] Mueller-Newen G., Stoffel W. Biol. Chem. Hoppe-Seyler 372:613-624(1991).

[3] Palosaari P.M., Hiltunen J.K. J. Biol. Chem. 265:2446-2449(1990).

[4] Nakahigashi K., Inokuchi H. Nucleic Acids Res. 18:4937-4937(1990).

[5] Driscoll J.R., Taber H.W. J. Bacteriol. 174:5063-5071(1992).

[6] Babbitt P.C., Kenyon G.L., Matin B.M., Charest H., Sylvestre M., Scholten J.D., Chang
 20 K.-H., Liang P.-H., Dunaway-Mariano D. Biochemistry 31:5594-5604(1992).

[7] Beckman D.L., Kranz R.G. Gene 107:171-172(1991).

[8] Eichler K., Bourgis F., Buchet A., Kleber H.-P., Mandrand-Berthelot M.-A. Mol. Microbiol. 13:775-786(1994).

25 166. (EF1BD) Elongation factor 1 beta/beta'/delta chain signatures

Eukaryotic elongation factor 1 (EF-1) is responsible for the GTP-dependent binding of aminoacyl-tRNAs to the ribosomes [1]. EF-1 is composed of four subunits: the alpha chain which binds GTP and aminoacyl-tRNAs, the gamma chain that probably plays a role in
 30 anchoring the complex to other cellular components and the beta and delta (or beta') chains. The beta and delta chains are highly similar proteins that both stimulate the exchange of GDP bound to the alpha chain for GTP [2]. The beta and delta chains are hydrophilic proteins of around 23 to 31 Kd. Their C-terminal part seems important for the nucleotide exchange

activity, while the N-terminal section is probably involved in the interaction with the gamma chain. Two signature patterns for this family of proteins were developed. The first corresponds to an acidic region in the central section; the second, to the C-terminal extremity of these proteins

5

Consensus pattern: [DE]-[DEG]-[DE](2)-[LIVMF SEQ ID NO:2)]-D-L-F-G-

Consensus pattern: [IV]-Q-S-x-D-[LIVM SEQ ID NO:4)]-x-A-[FWM]-[NQ]-K-[LIVM SEQ ID NO:4)]-

10 [1] Riis B., Rattan I.S., Clark B.F.C., Merrick W.C. Trends Biochem. Sci. 15:420-424(1990).

[2] van Damme H.T.F., Amons R., Karssies R., Timmers C.J., Janssen G.M.C., Moeller W. Biochim. Biophys. Acta 1050:241-247(1990).

15 167. (EF1G_domain) Elongation factor 1 gamma, conserved domain

168. (EFG_C) Elongation factor G C-terminus

20 This family is always found associated with GTP_EFTU. This family includes the carboxyl terminal regions of Elongation factor G, elongation factor 2 and some tetracycline resistance proteins.

169. (EFP) Elongation factor P signature

25 Elongation factor P (EF-P) [1] is a prokaryotic protein translation factor required for efficient peptide bond synthesis on 70S ribosomes from fMet-tRNA^{fMet}. EF-P is a protein of 21 Kd. It is evolutionary related to yeiP, an hypothetical protein from Escherichia coli. As a signature pattern, a conserved region located in the C-terminal part of these proteins was selected.

30

Consensus pattern: K-x-[AV]-x(4)-G-x(2)-[LIV]-x-V-P-x(2)-[LIV]-x(2)-G-

[1] Aoki H., Adams S.-L., Turner M.A., Ganoza M.C. Biochimie 79:7-11(1997).

170. (EF TS) Elongation factor Ts signatures

In prokaryotes elongation factor Ts (EF-Ts) is a component of the elongation cycle of protein biosynthesis. It associates with the EF-Tu.GDP complex and induces the exchange of GDP to GTP, it remains bound to the aminoacyl-tRNA.EF-Tu.GTP complex up to the GTP hydrolysis stage on the ribosome [1].EF-Ts is also a component of the chloroplast protein biosynthetic machinery and is encoded in the genome of some algal chloroplast [2]. It is also present in mitochondria [3]. As signature patterns for EF-Ts, two conserved regions located in the N-terminal part of the protein have been selected.

Consensus pattern: L-R-x(2)-T-[GSDNQ SEQ ID NO:231)]-x-[GS]-[LIVMF SEQ ID NO:2)]-x(0,1)-[DENKAC SEQ ID NO:232)]-x-K- [KRNEQS SEQ ID NO:233)]-A-L-
Consensus pattern: E-[LIVM SEQ ID NO:4)]-[NV]-[SCV]-[QE]-T-D-F-V-[SA]-[KRN]-

- [1] Bubunenko M.G., Kireeva M.L., Gudkov A.T. Biochimie 74:419-425(1992).
[2] Kostrzewa M., Zetsche K. Plant Mol. Biol. 23:67-76(1993).
[3] Xin H., Worlax V.L., Burkhardt W.A., Spremulli L.L. J. Biol. Chem. 270:17243-17249(1995).

171. (EMP24_GP25L) emp24/gp25L/p24 family

Members of this family are implicated in bringing cargo forward from the ER and binding to coat proteins by their cytoplasmic domains. Number of members: 30

Paccaud JP, Thomas DY, Bergeron JJ, Nilsson T, J Cell Biol 1998;140:751-765.

172. ENV_polyprotein

ENV polyprotein (coat polyprotein)

Number of members: 224

173. (ERG4_ERG24) Ergosterol biosynthesis ERG4/ERG24 family signatures

Two fungal enzymes involved in ergosterol biosynthesis and which act by reducing double bonds in precursors of ergosterol have been shown to be evolutionary related [1]. These are C-14 sterol reductase (gene ERG24 in budding yeast and *erg3* in *Neurospora Crassa*) and C-24(28) sterol reductase (gene ERG4 in budding yeast and *sts1* in fission yeast). Their sequences are also highly related to that of chicken lamin B receptor, which is thought to anchor the lamina to the inner nuclear membrane. These proteins are highly hydrophobic and seem to contain seven or eight transmembrane regions. As signature patterns, two conserved regions were selected. The first one is apparently located in a loop between the fourth and fifth transmembrane regions and the second is in the C-terminal section.

Consensus pattern: G-x(2)-[LIVM SEQ ID NO:4)]-[YH]-D-x-[FYW]-x-G-x(2)-L-N-P-R -
Consensus pattern: [LIVM SEQ ID NO:4)](2)-H-R-x(2)-R-D-x(3)-C-x(2)-K-Y-G-

[1] Lai M.H., Bard M., Pierson C.A., Alexander J.F., Goebel M., Carter G.T., Kirsch D.R.
Gene 140:41-49(1994).

174. (ERM) Ezrin/radixin/moesin family

This family of proteins contain a band 4.1 domain (Band_41), at their amino terminus.

This family represents the rest of these proteins.

[1] Yonemura S, Hirao M, Doi Y, Takahashi N, Kondo T, Tsukita S, J Cell Biol 1998;140:885-895.

175. ER lumen protein retaining receptor signatures

Proteins that reside in the lumen of the endoplasmic reticulum (ER) contain a C-terminal tetrapeptide (generally K-D-E-L or H-D-E-L) that serves as a signal for their retrieval (retrograde transport) from subsequent compartments of the secretory pathway. The signal is recognized by a receptor molecule that is believed to cycle between the cis side of the Golgi apparatus and the ER [1]. This protein is known as the ER lumen protein retaining receptor or also as the 'KDEL receptor'. It has been characterized in a variety of species, including fungi (gene ERD2), plants, Plasmodium, Drosophila and mammals. In mammals two highly related forms of the receptor are known. Structurally, the receptor is a protein of about 220 residues

that seems to contain seven transmembrane regions [2]. The N-terminal part (3 residues) is oriented toward the lumen while the C-terminal tail (about 12 residues) is cytoplasmic. There are three luminal and three cytoplasmic loops. Two signature patterns for these receptors were developed. The first pattern corresponds to the C-terminal half of the first cytoplasmic loop as well as most of the second transmembrane domain. The second pattern is a perfectly conserved decapeptide that corresponds to the central part of the fifth transmembrane domain.

Consensus pattern: G-I-S-x-[KR]-x-Q-x-L-[FY]-x-[LIV](2)-F-x(2)-R-Y-

Consensus pattern: L-E-[SA]-V-A-I-[LM]-P-Q-L-

[1] Pelham H.R.B. Curr. Opin. Cell Biol. 3:585-591(1991).

[2] Townsley F.M., Wilson D.W., Pelham H.R.B. EMBO J. 12:2821-2829(1993).

176. (ETF_beta) Electron transfer flavoprotein beta-subunit signature

The electron transfer flavoprotein (ETF) [1,2] serves as a specific electron acceptor for various mitochondrial dehydrogenases. ETF transfers electrons to the main respiratory chain via ETF-ubiquinone oxidoreductase. ETF is an heterodimer that consist of an alpha and a beta subunit and which bind one molecule of FAD per dimer. A similar system also exists in some bacteria. The beta subunit of ETF is a protein of about 28 Kd which is structurally related to the bacterial nitrogen fixation protein fixA which could play a role in a redox process and feed electrons to ferredoxin. Other related proteins are: - Escherichia coli hypothetical protein ydiQ. - Escherichia coli hypothetical protein ygcR. As a signature pattern for these proteins, a conserved region which is located in the central section was selected.

Consensus pattern: [IVA]-x-[KR]-x(2)-[DE]-[GD]-[GDE]-x(1,2)-[EQ]-x-[LIV]- x(4)-P-x-[LIVM SEQ ID NO:4)](2)-[TAC]-

[1] Finocchiaro G., Ikeda Y., Ito M., Tanaka K. Prog. Clin. Biol. Res. 321:637-652(1990).

[2] Tsai M.H., Saier M.H. Jr. Res. Microbiol. 146:397-404(1995).

177. Endonuclease III signatures

Escherichia coli endonuclease III (EC 4.2.99.18) (gene nth) [1] is a DNA repair enzyme that acts both as a DNA N-glycosylase, removing oxidized pyrimidines from DNA, and as an apurinic/apyrimidinic (AP) endonuclease, introducing a single-strand nick at the site from which the damaged base was removed. Endonuclease III is an iron-sulfur protein that binds a single 4Fe-4S cluster. The 4Fe-4S cluster does not seem to be important for catalytic activity, but is probably involved in the proper positioning of the enzyme along the DNA strand [2]. Endonuclease III is evolutionary related to the following proteins: - Fission yeast endonuclease III homolog (gene nth1) [3]. - Escherichia coli and related protein DNA repair protein mutY, which is an adenine glycosylase. MutY is a larger protein (350 amino acids) than endonuclease III (211 amino acids). - Micrococcus luteus ultraviolet N-glycosylase/AP lyase which initiates repair at cis-syn pyrimidine dimers. - ORF10 in plasmid pFV1 of the thermophilic archaebacteria Methanobacterium thermoformicum [4]. Restriction methylase m.MthTI, which is encoded by this plasmid, generates 5-methylcytosine which is subject to deamination resulting in G-T mismatches. This protein could correct these mismatches. - Yeast hypothetical protein YAL015c. - Fission yeast hypothetical protein SpAC26A3.02. - Caenorhabditis elegans hypothetical protein R10E4.5. - Methanococcus jannaschii hypothetical protein MJ0613. The 4Fe-4S cluster is bound by four cysteines which are all located in a 17 amino acid region at the C-terminal end of endonuclease III. A similar region is also present in the central section of mutY and in the C-terminus of ORF10 and of the Micrococcus UV endonuclease. The 4Fe-4S cluster region does not exist in YAL015c. Two signature patterns for these proteins were developed: the first corresponds to the core of the iron-sulfur binding domain, the second corresponds to the best conserved region in the catalytic core of these enzymes.

Consensus pattern: C-x(3)-[KRS]-P-[KRAGL SEQ ID NO:234]-C-x(2)-C-x(5)-C [The four C's are 4Fe-4S ligands]-

Consensus pattern: [GST]-x-[LIVMF SEQ ID NO:2]-P-x(5)-[LIVMW SEQ ID NO:235])-x(2,3)-[LI]-[PAS]-G-V-[GA]-x(3)-[GAC]-x(3)-[LIVM SEQ ID NO:4])-x(2)-[SALV SEQ ID NO:236])-[LIVMFYW SEQ ID NO:26])-[GANK SEQ ID NO:237])-

[1] Kuo C.-F., McRee D., Fisher C.L., O'Handley S.F., Cunningham R.P., Tainer J.A. Science 258:434-440(1992).

[2] Thomson A.J. Curr. Biol. 3:173-174(1993).

[3] Roldan-Arjona T., Anselmino C., Lindahl T. Nucleic Acids. Res. 3307-3312(1996).

[4] Noelling J., van Eeden F.J.M., Eggen R.I.L., de Vos W.M. Nucleic Acids Res. 20:6501-6507(1992).

5

178. (Epimerase) NAD dependent epimerase/dehydratase family

This family of proteins utilize NAD as a cofactor. The proteins in this family use nucleotide-sugar substrates for a variety of chemical reactions.

10

[1] Thoden JB, Hegeman AD, Wesenberg G, Chapeau MC, Frey PA, Holden HM, Biochemistry 1997;36:6294-6304.

179. Exonuclease

15

This family includes a variety of exonuclease proteins, such as ribonuclease T and the epsilon subunit of DNA polymerase III.

[1] Koonin EV, Deutscher MP, Nucleic Acids Res 1993;21:2521-2522.

20

180. ENTH

ENTH domain

[1] Kay BK, Yamabhai M, Wendland B, Emr SD; Medline: 99156083, Identification of a novel domain shared by putative components of the endocytic and cytoskeletal machinery.

25

Protein Sci 1999;8:435-438.

The ENTH (Epsin N-terminal homology) domain is found in proteins involved in endocytosis and cytoskeletal machinery. The function of the ENTH domain is unknown.

30

Number of members: 29

181. (eIF-1A) Eukaryotic initiation factor 1A signature

211

Eukaryotic translation initiation factor 1A (eIF-1A) [1] (formerly known as eIF-4C) is a protein that seems to be required for maximal rate of protein biosynthesis. It enhances ribosome dissociation into subunits and stabilizes the binding of the initiator Met-tRNA to 40S ribosomal subunits. eIF-1A is a hydrophilic protein of about 15 to 17 Kd. Archaeobacteria also seem to possess a eIF-1A homolog. As a signature pattern, a conserved region in the central section of these proteins was selected.

Consensus pattern: [IM]-x-G-x-[GS]-[KRH]-x(4)-[CL]-x-D-G-x(2)-R-x(2)-[RH]-I-x-G

[1] Wei C.-L., Kainuma M., Hershey J.W.B. *J. Biol. Chem.* 270:22788-22794(1995).

182. (eIF-5A) Eukaryotic initiation factor 5A hypusine signature

Eukaryotic initiation factor 5A (eIF-5A) (formerly known as eIF-4D) [1,2] is a small protein whose precise role in the initiation of protein synthesis is not known. It appears to promote the formation of the first peptide bond. eIF-5A seems to be the only eukaryotic protein to contain an hypusine residue. Hypusine is derived from lysine by the post-translational addition of a butylamino group (from spermidine) to the epsilon-amino group of lysine. The hypusine group is essential to the function of eIF-5A. A hypusine-containing protein has been found in archaeobacteria such as *Sulfolobus acidocaldarius* or *Methanococcus jannaschii*; this protein is highly similar to eIF-5A and could play a similar role in protein biosynthesis. The signature developed for eIF-5A is centered around the hypusine residue.

Consensus pattern: [PT]-G-K-H-G-x-A-K [The first K is modified to hypusine]

[1] Park M.H., Wolff E.C., Folk J.E. *Biofactors* 4:95-104(1993).

[2] Schnier J., Schwelberger H.G., Smit-McBride Z., Kang H.A., Hershey J.W.B. *Mol. Cell. Biol.* 11:3105-3114(1991).

183. (efhand) S-100/ICaBP type calcium binding protein signature

S-100 are small dimeric acidic calcium and zinc-binding proteins [1] abundant in the brain. They have two different types of calcium-binding sites: a low affinity one with a special

structure and a 'normal' EF-hand type high affinity site. The vitamin-D dependent intestinal calcium-binding proteins (ICaBP or calbindin 9 Kd) also belong to this family of proteins, but it does not form dimers. In the past years the sequences of many new members of this family have been determined (for reviews see [2,3,4]); in most cases the function of these proteins is not yet known, although it is becoming clear that they are involved in cell growth and differentiation, cell cycle regulation and metabolic control. These proteins are: - Calcyclin (Prolactin receptor associated protein (PRA); clatropin; 2a9; 5B10; S100A6). - Calpactin I light chain (p10; p11; 42c; S100A10). - Calgranulin A (cystic fibrosis antigen (CFAg); MIF related protein 8 (MRP- 8); p8; S100A8). - Calgranulin B (MIF related protein 14 (MRP-14); p14; S100A9). - Calgizzarin (S100C). - Placental calcium-binding protein (CAPL) (18a2; peL98; 42a; p9K; MTS1; metastatin; S100A4). - Protein S-100D (S100A5). - Protein S-100E (S100A3). - Protein S-100L (CAN19; S100A2). - Placental protein S-100P (S100E). - Psoriasin (S100A7). - Chemotactic cytokine CP-10 [5]. - Protein MRP-126 [6]. - Trichohyalin [7]. This is a large intermediate filament-associated protein that associates with keratin intermediate filaments (KIF); it contains a S- 100 type domain in its N-terminal extremity. A number of these proteins are known to bind calcium while others are not (p10 for example). Our EF-hand detecting pattern will fail to pick those proteins which have lost their calcium-binding properties. A pattern was developed which unambiguously picks up proteins belonging to this family. This pattern spans the region of the EF-hand high affinity site but makes no assumptions on the calcium-binding properties of this site.

Consensus pattern: [LIVMFYW SEQ ID NO:26)](2)-x(2)-[LK]-D-x(3)-[DN]-x(3)-[DNSG SEQ ID NO:238)]-[FY]-x- [ES]-[FYVC SEQ ID NO:239)]-x(2)-[LIVMFS SEQ ID NO:132)]-[LIVMF SEQ ID NO:2)]

[1] Baudier J. (In) Calcium and Calcium Binding proteins, Gerday C., Bollis L., Giller R., Eds., pp102-113, Springer Verlag, Berlin, (1988).

[2] Moncrief N.D., Kretsinger R.H., Goodman M. J. Mol. Evol. 30:522-562(1990).

[3] Kligman D., Hilt D.C. Trends Biochem. Sci. 13:437-443(1988).

[4] Schaefer B.W., Wicki R., Engelkamp D., Mattei M.-G., Heizmann C.W. Genomics 25:638-643(1995).

[5] Lackmann M., Cornish C.J., Simpson R.J., Moritz R.L., Geczy C.L. J. Biol. Chem. 267:7499-7504(1992).

[6] Nakano T., Graf T. Oncogene 7:527-534(1992).

[7] Lee S.-C., Kim I.-G., Marekov L.N., O'Keefe E.J., Parry D.A.D., Steinert P.M., J. Biol. Chem. 268:12164-12176(1993).

EF-hand calcium-binding domain

Many calcium-binding proteins belong to the same evolutionary family and share a type of calcium-binding domain known as the EF-hand [1 to 5]. This type of domain consists of a twelve residue loop flanked on both side by a twelve residue alpha-helical domain. In an EF-hand loop the calcium ion is coordinated in a pentagonal bipyramidal configuration. The six residues involved in the binding are in positions 1, 3, 5, 7, 9 and 12; these residues are denoted by X, Y, Z, -Y, -X and -Z. The invariant Glu or Asp at position 12 provides two oxygens for liganding Ca (bidentate ligand).

Listed below are the proteins which are known to contain EF-hand regions. For each type of protein the total number of EF-hand regions known or supposed to exist is indicated between parenthesis. This number does not include regions which clearly have lost their calcium-binding properties, or the atypical low-affinity site (which spans thirteen residues) found in the S-100/ICaBP family of proteins [6].

- Aequorin and Renilla luciferin binding protein (LBP) (Ca=3).
- Alpha actinin (Ca=2). - Calbindin (Ca=4).
- Calcineurin B subunit (protein phosphatase 2B regulatory subunit) (Ca=4).
- Calcium-binding protein from *Streptomyces erythraeus* (Ca=3?).
- Calcium-binding protein from *Schistosoma mansoni* (Ca=2?).
- Calcium-binding proteins TCBP-23 and TCBP-25 from *Tetrahymena thermophila* (Ca=4?). - Calcium-dependent protein kinases (CDPK) from plants (Ca=4).
- Calcium vector protein from *amphoxius* (Ca=2).
- Calcyphosin (thyroid protein p24) (Ca=4?).
- Calmodulin (Ca=4, except in yeast where Ca=3).
- Calpain small and large chains (Ca=2). - Calretinin (Ca=6).
- Calcyclin (prolactin receptor associated protein) (Ca=2).

- Caltractin (centrin) (Ca=2 or 4).
 - Cell Division Control protein 31 (gene CDC31) from yeast (Ca=2?).
 - Diacylglycerol kinase (EC 2.7.1.107) (DGK) (Ca=2).
 - FAD-dependent glycerol-3-phosphate dehydrogenase (EC 1.1.99.5) from mammals (Ca=1).
 - 5 - Fimbrin (plastin) (Ca=2).
 - Flagellar calcium-binding protein (1f8) from *Trypanosoma cruzi* (Ca=1 or 2).
 - Guanylate cyclase activating protein (GCAP) (Ca=3).
 - Inositol phospholipid-specific phospholipase C isozymes gamma-1 and delta-1 (Ca=2) [10].
 - Intestinal calcium-binding protein (ICaBPs) (Ca=2).
 - 10 - MIF related proteins 8 (MRP-8 or CFAG) and 14 (MRP-14) (Ca=2).
 - Myosin regulatory light chains (Ca=1).
 - Oncomodulin (Ca=2).
 - Osteonectin (basement membrane protein BM-40) (SPARC) and proteins that contains an 'osteonectin' domain (QR1, matrix glycoprotein SC1) (see the entry <PDOC00535>) (Ca=1).
 - Parvalbumins alpha and beta (Ca=2).
 - 15 - Placental calcium-binding protein (18a2) (nerve growth factor induced protein 42a) (p9k) (Ca=2).
 - Recoverins (visinin, hippocalcin, neurocalcin, S-modulin) (Ca=2 to 3).
 - Reticulocalbin (Ca=4).
 - S-100 protein, alpha and beta chains (Ca=2).
 - Sarcoplasmic calcium-binding protein (SCPs) (Ca=2 to 3).
 - 20 - Sea urchin proteins Spec 1 (Ca=4), Spec 2 (Ca=4?), Lps-1 (Ca=8).
 - Serine/threonine protein phosphatase rdgc (EC 3.1.3.16) from *Drosophila* (Ca=2).
 - Sorcin V19 from hamster (Ca=2).
 - Spectrin alpha chain (Ca=2).
 - Squidulin (optic lobe calcium-binding protein) from squid (Ca=4).
 - Troponins C; from skeletal muscle (Ca=4), from cardiac muscle (Ca=3), from arthropods and molluscs (Ca=2).
 - 25
- There has been a number of attempts [7,8] to develop patterns that pick-up EF-hand regions, but these studies were made a few years ago when not so many different families of calcium-binding proteins were known. Therefore a new pattern was developed which takes into account all published sequences. This
- 30 pattern includes the complete EF-hand loop as well as the first residue which follows the loop and which seem to always be hydrophobic.

-Consensus pattern: D-x-[DNS]-{ILVFIYW SEQ ID NO:240}}-[DENSTG SEQ ID NO:241)]-[DNQGHRK SEQ ID NO:242)]-{GP}-[LIVMC SEQ ID NO:142)]-[DENQSTAGC SEQ ID NO:243)]-x(2)-[DE]-[LIVMFYW SEQ ID NO:26)]

-Note: positions 1 (X), 3 (Y) and 12 (-Z) are the most conserved.

5 -Note: the 6th residue in an EF-hand loop is, in most cases a Gly, but the number of exceptions to this 'rule' has gradually increased and therefore the pattern should include all the different residues which have been shown to exist in this position in functional Ca-binding sites.

10 -Note: the pattern will, in some cases, miss one of the EF-hand regions in some proteins with multiple EF-hand domains.

[1] Kawasaki H., Kretsinger R.H. Protein Prof. 2:305-490(1995).[2] Kretsinger R.H. Cold Spring Harbor Symp. Quant. Biol. 52:499-510(1987).

[3] Moncrief N.D., Kretsinger R.H., Goodman M. J. Mol. Evol. 30:522-562(1990).

15 [4] Nakayama S., Moncrief N.D., Kretsinger R.H. J. Mol. Evol. 34:416-448(1992).

[5] Heizmann C.W., Hunziker W. Trends Biochem. Sci. 16:98-103(1991).

[6] Kligman D., Hilt D.C. Trends Biochem. Sci. 13:437-443(1988).

[7] Strynadka N.C.J., James M.N.G.

Annu. Rev. Biochem. 58:951-98(1989).

20 [8] Haiech J., Sallantin J. Biochimie 67:555-560(1985).

[9] Chauvaux S., Beguin P., Aubert J.-P., Bhat K.M., Gow L.A., Wood T.M., Bairoch A. Biochem. J. 265:261-265(1990).

[10] Bairoch A., Cox J.A. FEBS Lett. 269:454-456(1990).

25

184. Enolase signature

Enolase (EC 4.2.1.11) is a glycolytic enzyme that catalyzes the dehydration of 2-phospho-D-glycerate to phosphoenolpyruvate [1]. It is a dimeric enzyme that requires magnesium both for catalysis and stabilizing the dimer. Enolase is probably found in all organisms that
30 metabolize sugars. In vertebrates, there are three different tissue-specific isozymes: alpha present in most tissues, beta in muscles and gamma found only in nervous tissues. Tau-crystallin, one of the major lens proteins in some fish, reptiles and birds, has been shown [2]

to be evolutionary related to enolase. As a signature pattern for enolase, the best conserved region was selected, it is located in the C-terminal third of the sequence.-

Consensus pattern: [LIV](3)-K-x-N-Q-I-G-[ST]-[LIV]-[ST]-[DE]-[STA]

5 [1] Lebioda L., Stec B., Brewer J.M. J. Biol. Chem. 264:3685-3693(1989).

[2] Wistow G., Piattigorsky J. Science 236:1554-1556(1987).

185. (F-actin_cap_A) F-actin capping protein alpha subunit signatures

10 The F-actin capping protein binds in a calcium-independent manner to the fast growing ends of actin filaments (barbed end) thereby blocking the exchange of subunits at these ends. Unlike gelsolin and severin this protein does not sever actin filaments. The F-actin capping protein is a heterodimer composed of two unrelated subunits: alpha and beta. The alpha subunit is a protein of about 268 to 286 amino acid residues whose sequence is well
15 conserved in eukaryotic species [1]. As signature patterns two highly conserved regions in the C-terminal section of the alpha subunit were selected.

Consensus pattern: V-H-[FY](2)-E-D-G-N-V

Consensus pattern: F-K-[AE]-L-R-R-x-L-P-

20

[1] Cooper J.A., Caldwell J.E., Gattermeir D.J., Torres M.A., Amatruda J.F., Casella J.F. Cell Motil. Cytoskeleton 18:204-214(1991).

25 186. F-box domain

[1] Bai C, Sen P, Hofmann K, Ma L, Goebel M, Harper JW, Elledge SJ, Cell 1996;86:263-274. [2] Skowyra D, Craig KL, Tyers M, Elledge SJ, Harper JW, Cell 1997;91:209-219.

30

187. F-protein

Negative factor, (F-Protein) or Nef.

[1] Arold S, Franken P, Strub M-P, Hoh F, Benichou S, Benarous R, Dumas C; Medline: 98035457, The crystal structure of HIV-1 Nef protein bound to the Fyn kinase SH3 domain suggests a role for this complex in altered T cell receptor signalling Structure 1997;5:1361-1372.

5

Nef protein accelerates virulent progression of AIDS by its interaction with cellular proteins involved in signal transduction and host cell activation. Nef has been shown to bind specifically to a subset of the Src kinase family.

10 Number of members: 1013

188. (FAD_binding_2)

Fumarate reductase / succinate dehydrogenase FAD-binding site

15

In bacteria two distinct, membrane-bound, enzyme complexes are responsible for the interconversion of fumarate and succinate (EC 1.3.99.1): fumarate reductase (Frd) is used in anaerobic growth, and succinate dehydrogenase (Sdh) is used in aerobic growth. Both complexes consist of two main components: a membrane-extrinsic component composed of a FAD-binding flavoprotein and an iron-sulfur protein; and an hydrophobic component composed of a membrane anchor protein and/or a cytochrome B.

20

In eukaryotes mitochondrial succinate dehydrogenase (ubiquinone) (EC 1.3.5.1) is an enzyme composed of two subunits: a FAD flavoprotein and an iron-sulfur protein.

25

The flavoprotein subunit is a protein of about 60 to 70 Kd to which FAD is covalently bound to a histidine residue which is located in the N-terminal section of the protein [1]. The sequence around that histidine is well conserved in Frd and Sdh from various bacterial and eukaryotic species [2] and can be used as a signature pattern.

30

Consensus pattern R-[ST]-H-[ST]-x(2)-A-x-G-G [H is the FAD binding site] Sequences known to belong to this class detected by the pattern ALL.

[1] Blaut M., Whittaker K., Valdovinos A., Ackrell B.A., Gunsalus R.P., Cecchini G. J. Biol. Chem. 264:13599-13604(1989).

[2] Birch-Machin M.A., Farnsworth L., Ackrell B.A., Cochran B., Jackson S., Bindoff L.A., Aitken A., Diamond A.G., Turnbull D.M. J. Biol. Chem. 267:11553-11558(1992).

5

189. Fatty acid desaturases signatures (FA_desaturase)

Fatty acid desaturases (EC 1.14.99.-) are enzymes that catalyze the insertion of a double bond at the delta position of fatty acids. There seems to be two distinct families of fatty acid

10 desaturases which do not seem to be evolutionary related. Family 1 is composed of: -

Stearoyl-CoA desaturase (SCD) (EC 1.14.99.5) [1]. SCD is a key regulatory enzyme of unsaturated fatty acid biosynthesis. SCD introduces a cis double bond at the delta(9) position of fatty acyl-CoA's such as palmitoleoyl- and oleoyl-CoA. SCD is a membrane-bound enzyme that is thought to function as a part of a multienzyme complex in the endoplasmic

15 reticulum of vertebrates and fungi. As a signature pattern for this family a conserved region in the C-terminal part of these enzymes was selected, this region is rich in histidine residues and in aromatic residues. Family 2 is composed of: - Plants stearoyl-acyl-carrier-protein

desaturase (EC 1.14.99.6) [2], these enzymes catalyze the introduction of a double bond at the delta(9) position of stearoyl-ACP to produce oleoyl-ACP. This enzyme is responsible for

20 the conversion of saturated fatty acids to unsaturated fatty acids in the synthesis of vegetable oils. - Cyanobacteria desA [3] an enzyme that can introduce a second cis double bond at the

delta(12) position of fatty acid bound to membranes glycerolipids. DesA is involved in chilling tolerance; the phase transition temperature of lipids of cellular membranes being

25 dependent on the degree of unsaturation of fatty acids of the membrane lipids. As a signature pattern for this family a conserved region in the C-terminal part of these enzymes was selected.

Consensus pattern: G-E-x-[FY]-H-N-[FY]-H-H-x-F-P-x-D-Y-

Consensus pattern: [ST]-[SA]-x(3)-[QR]-[LI]-x(5,6)-D-Y-x(2)-[LIVMFYW SEQ ID

30 NO:26)]-[LIVM SEQ ID NO:4)]- [DE]-

[1] Kaestner K.H., Ntambi J.M., Kelly T.J. Jr., Lane M.D. J. Biol. Chem. 264:14755-14761(1989).

[2] Shanklin J., Somerville C.R. Proc. Natl. Acad. Sci. U.S.A. 88:2510-2514(1991).

[3] Wada H., Gombos Z., Murata N. Nature 347:200-203(1990).

5 190. Fructose-1-6-bisphosphatase active site (FBPase)

Fructose-1,6-bisphosphatase (EC 3.1.3.11) (FBPase) [1], a regulatory enzyme in gluconeogenesis, catalyzes the hydrolysis of fructose 1,6-bisphosphate to fructose 6-phosphate. It is involved in many different metabolic pathways and found in most organisms. Sedoheptulose-1,7-bisphosphatase (EC 3.1.3.37) (SBPase) [2] is an enzyme found
10 plant chloroplast and in photosynthetic bacteria that catalyzes the hydrolysis of sedoheptulose 1,7-bisphosphate to sedoheptulose 7-phosphate, a step in the Calvin's reductive pentose phosphate cycle. It is functionally and structurally related to FBPase. In mammalian FBPase, a lysine residue has been shown to be involved in the catalytic mechanism [3]. The region around this residue is highly conserved and can be used as a signature pattern for FBPase and
15 SBPase. It must be noted that, in some bacterial FBPase sequences, the active site lysine is replaced by an arginine

Consensus pattern: [AG]-[RK]-L-x(1,2)-[LIV]-[FY]-E-x(2)-P-[LIVM SEQ ID NO:4)]-[GSA]
[K/R is the active site residue]-

20

[1] Benkovic S.J., DeMaine M.M. Adv. Enzymol. 53:45-82(1982).

[2] Raines C.A., Lloyd J.C., Willingham N.M., Potts S., Dyer T.A. Eur. J. Biochem. 205:1053-1059(1992).

[3] Ke H., Thorpe C.M., Seaton B.A., Lipscomb W.N., Marcus F. J. Mol. Biol. 212:513-
25 539(1989).

191. FGGY family of carbohydrate kinases signatures *

It has been shown [1] that four different type of carbohydrate kinases seem to be evolutionary
30 related. These enzymes are: - L-fucolokinase (EC 2.7.1.51) (gene fucK). - Gluconokinase (EC 2.7.1.12) (gene gntK). - Glycerokinase (EC 2.7.1.30) (gene glpK). - Xylulokinase (EC 2.7.1.17) (gene xylB). - L-xylulose kinase (EC 2.7.1.53) (gene lyxK). These enzymes are proteins of from 480 to 520 amino acid residues. As consensus patterns for this family of

kinases two conserved regions were selected, one in the central section, the other in the C-terminal section.

Consensus pattern: [MFYGS SEQ ID NO:244)]-x-[PST]-x(2)-K-[LIVMFYW SEQ ID NO:26)]-x-W-[LIVMF SEQ ID NO:2)]-x-[DENQTKR SEQ ID NO:245)]- [ENQH SEQ ID NO:246)]-

Consensus pattern: [GSA]-x-[LIVMFYW SEQ ID NO:26)]-x-G-[LIVM SEQ ID NO:4)]-x(7,8)-[HDENQ SEQ ID NO:247)]-[LIVMF SEQ ID NO:2)]-x(2)- [AS]-[STAIVM SEQ ID NO:248)]-[LIVMFY SEQ ID NO:18)]-[DEQ]-

[1] Reizer A., Deutscher J., Saier M.H. Jr., Reizer J. Mol. Microbiol. 5:1081-1089(1991).

192. FKBP-type peptidyl-prolyl cis-trans isomerase signatures/profile (FKBP)

FKBP [1,2,3] is the major high-affinity binding protein, in vertebrates, for the immunosuppressive drug FK506. It exhibits peptidyl-prolyl cis-trans isomerase activity (EC 5.2.1.8) (PPIase or rotamase). PPIase is an enzyme that accelerates protein folding by catalyzing the cis-trans isomerization of proline imidic peptide bonds in oligopeptides [4]. At least three different forms of FKBP are known in mammalian species: - FKBP-12, which is cytosolic and inhibited by both FK506 and rapamycin. - FKBP-13, which is membrane associated and inhibited by both FK506 and rapamycin. - FKBP-25, which is preferentially inhibited by rapamycin. These forms of FKBP are evolutionary related and show extensive similarities[5,6,7] with the following proteins: - Fungal FKBP. - Mammalian hsp binding immunophilin (HBI) (also called p59). HBI is a protein which binds to hsp90 and contains two FKBP-like domains in its N- terminal section - the first of which seems to be functional. - The C-terminal part of the cell-surface protein mip from Legionella; a protein associated with macrophage infection by an unknown mechanism. - Escherichia coli slyD [8], a protein with a N-terminal FKBP domain followed by an histidine-rich metal-binding domain. - Escherichia coli fkpA. - Escherichia coli fklB (FKBP22). - Escherichia coli slpA. - Bacterial trigger factor (Tig). - Streptomyces hygroscopus and chrysomallus FK506-binding protein. - Chlamydia trachomatis 27 Kd membrane protein. - Neisseria meningitidis strain C114 PPIase. - Probable PPIases from Haemophilus influenzae (HI0754), Methanococcus jannaschii (MJ0278 and MJ0825), Pseudomonas fluorescens and Pseudomonas aeruginosa.

Two signature patterns for these proteins were developed. One is based on a conserved region in the N-terminus of FKBP, the other is located in the central section. The profile for FKBP spans the complete domain.

5 Consensus pattern: [LIVMC SEQ ID NO:142)]-x-[YF]-x-[GVL]-x(1,2)-[LFT]-x(2)-G-x(3)-[DE]- [STAEQK SEQ ID NO:249)]-[STAN SEQ ID NO:250)]-

Consensus pattern: [LIVMFY SEQ ID NO:18)]-x(2)-[GA]-x(3,4)-[LIVMF SEQ ID NO:2)]-x(2)-[LIVMFHK SEQ ID NO:251)]-x(2)-G- x(4)-[LIVMF SEQ ID NO:2)]-x(3)-[PSGAQ SEQ ID NO:252)]-x(2)-[AG]-[FY]-G--

10 [1] Tropschug M., Wachter E., Mayer S., Schoenbrunner E.R., Schmid F.X. Nature 346:674-677(1990).

[2] Stein R.L. Curr. Biol. 1:234-236(1991).

[3] Siekierka J.J., Widerrecht G., Greulich H., Boulton D., Hung S.H.Y., Cryan J., Hodges P.J., Sigal N.H. J. Biol. Chem. 265:21011-21015(1990).

[4] Fischer G., Schmid F.X. Biochemistry 29:2205-2212(1990).

[5] Trandinh C.C., Pao G.M., Saier M.H. Jr. FASEB J. 6:3410-3420(1992).

[6] Galat A. Eur. J. Biochem. 216:689-707(1993).

[7] Hacker J., Fischer G. Mol. Microbiol. 10:445-456(1993).

20 [8] Wuelfing C., Lomardero J., Plueckthun A. J. Biol. Chem. 269:2895-2901(1994).

193. MAPEG family (aka: FLAP/GST2/LTC4S family signature)

The following mammalian proteins are evolutionary related [1]:

- 25 - Leukotriene C4 synthase (EC 2.5.1.37) (gene LTC4S), an enzyme that catalyzes the production of LTC4 from LTA4.
- Microsomal glutathione S-transferase II (EC 2.5.1.18) (GST-II) (gene GST2), an enzyme that can also produces LTC4 from LTA4.
- 30 - 5-lipoxygenase activating protein (gene FLAP), a protein that seems to be required for the activation of 5-lipoxygenase.

These are proteins of 150 to 160 residues that contain three transmembrane segments. As a signature pattern, a conserved region between the first and second transmembrane domains was selected.

Consensus pattern: G-x(3)-F-E-R-V-[FY]-x-A-[NQ]-x-N-C

[1] Jakobsson P.-J., Mancini J.A., Ford-Hutchinson A.W. J. Biol. Chem. 271:22203-22210(1996).

194. FMN-dependent alpha-hydroxy acid dehydrogenases active site (FMN_dh)

A number of oxidoreductases that act on alpha-hydroxy acids and which are FMN-containing flavoproteins have been shown [1,2,3] to be structurally related; these enzymes are: - Lactate dehydrogenase (EC 1.1.2.3), which consists of a dehydrogenase domain and a heme-binding domain called cytochrome b2 and which catalyzes the conversion of lactate into pyruvate. - Glycolate oxidase (EC 1.1.3.15) ((S)-2-hydroxy-acid oxidase), a peroxisomal enzyme that catalyzes the conversion of glycolate and oxygen to glyoxylate and hydrogen peroxide. - Long chain alpha-hydroxy acid oxidase from rat (EC 1.1.3.15), a peroxisomal enzyme. - Lactate 2-monooxygenase (EC 1.13.12.4) (lactate oxidase) from *Mycobacterium smegmatis*, which catalyzes the conversion of lactate and oxygen to acetate, carbon dioxide and water. - (S)-mandelate dehydrogenase from *Pseudomonas putida* (gene mdlB), which catalyzes the reduction of (S)-mandelate to benzoylformate. The first step in the reaction mechanism of these enzymes is the abstraction of the proton from the alpha-carbon of the substrate producing a carbanion which can subsequently attach to the N5 atom of FMN. A conserved histidine has been shown [4] to be involved in the removal of the proton. The region around this active site residue is highly conserved and contains an arginine residue which is involved in substrate binding.

Consensus pattern: S-N-H-G-[AG]-R-Q [H is the active site residue] [R is a substrate-binding residue]-

[1] Giegel D.A., Williams C.H. Jr., Massey V. J. Biol. Chem. 265:6626-6632(1990).

[2] Tsou A.Y., Ransom S.C., Gerlt J.A., Buechter D.D., Babbitt P.C., Kenyon G.L.

Biochemistry 29:9856-9862(1990).

[3] Le K.H.D., Lederer F. J. Biol. Chem. 266:20877-20880(1991).

[4] Lindqvist Y., Branden C.-I. J. Biol. Chem. 264:3624-3628(1989).

195. Flavin-binding monooxygenase-like (FMO-like)

This family includes FMO proteins, cyclohexanone monooxygenase

5

196. (FPGS)

Folylpolyglutamate synthase signatures (aka Mur_ligase)

10

Folylpolyglutamate synthase (EC 6.3.2.17) (FPGS) [1] is the enzyme of folate metabolism that catalyzes ATP-dependent addition of glutamate moieties to tetrahydrofolate.

Its sequence is moderately conserved between prokaryotes (gene folC) and eukaryotes. We developed two signature patterns based on the conserved regions which are rich in glycine residues and could play a role in the catalytical activity and/or in substrate binding.

15

Consensus pattern [LIVMFY SEQ ID NO:18])-x-[LIVM SEQ ID NO:4)]-[STAG SEQ ID NO:20)]-G-T-[NK]-G-K-x-[ST]-x(7)- [LIVM SEQ ID NO:4)](2)-x(3)-[GSK] Sequences known to belong to this class detected by the pattern ALL.

20

Consensus pattern[LIVMFY SEQ ID NO:18)](2)-E-x-G-[LIVM SEQ ID NO:4)]-[GA]-G-x(2)-D-x-[GST]-x-[LIVM SEQ ID NO:4)](2) Sequences known to belong to this class detected by the pattern ALL.

25

[1] Shane B., Garrow T., Brenner A., Chen L., Choi Y.J., Hsu J.C., Stover P. Adv. Exp. Med. Biol. 338:629-634(1993).

197. FYVE zinc finger

30

The FYVE zinc finger is named after four proteins that it has been found in: Fab1, YOTB/ZK632.12, Vac1, and EEA1. The FYVE finger has been shown to bind two Zn⁺⁺ ions [1]. The FYVE finger has eight potential zinc coordinating cysteine positions. Many members of this family also include two histidines in a motif R+HHC+XCG, where +

represents a charged residue and X any residue. Members were included which do not conserve these histidine residues but are clearly related.

[1] Stenmark H, Aasland R, Toh BH, D'Arrigo A, J Biol Chem 1996;271:24048-24054. [2] Gaullier JM, Simonsen A, D'Arrigo A, Bremnes B, Stenmark H, Aasland R, Nature 1998;394:432-433.

198. F_actin_cap_B

F-actin capping protein beta subunit signature

The F-actin capping protein binds in a calcium-independent manner to the fast growing ends of actin filaments (barbed end) thereby blocking the exchange of subunits at these ends. Unlike gelsolin and severin this protein does not sever actin filaments. The F-actin capping protein is a heterodimer composed of two unrelated subunits: alpha and beta.

The beta subunit is a protein of about 280 amino acid residues whose sequence is well conserved in eukaryotic species [1]. As a signature pattern a conserved hexapeptide in the N-terminal section of the beta subunit was selected.

Consensus pattern: C-D-Y-N-R-D Sequences known to belong to this class detected by the pattern ALL.

[1] Amatruda J.F., Cannon J.F., Tatchell K., Hug C., Cooper J.A. Nature 344:352-354(1990).

199. Isopenicillin N synthetase signatures (Fe_Asc_oxidored)

Isopenicillin N synthetase (IPNS) [1,2] is a key enzyme in the biosynthesis of penicillin and cephalosporin. In the presence of oxygen, it removes iron and ascorbate, four hydrogen atoms from L-(alpha-aminoadipyl)-L-cysteinyl-d-valine to form the azetidinone and thiazolidine rings of isopenicillin. IPNS is an enzyme of about 330 amino-acid residues. Two cysteines are conserved in fungal and bacterial IPNS sequences; these may be involved in iron-binding and/or substrate-binding. Cephalosporium acremonium DAOCS/DACS [3] is a bifunctional enzyme involved in cephalosporin biosynthesis. The DAOCS domain, which is structurally

related to IPNS, catalyzes the step from penicillin N to deacetoxy-cephalosporin C - used as a substrate by DACS to form deacetylcephalosporin C. *Streptomyces clavuligerus* possesses a monofunctional DAOCS enzyme (gene *cefE*) [4] also related to IPNS. Two signature patterns for these enzymes were derived, centered around the conserved cysteine residues.

5

Consensus pattern: [RK]-x-[STA]-x(2)-S-x-C-Y-[SL]-

Consensus pattern: [LIVM SEQ ID NO:4)](2)-x-C-G-[STA]-x(2)-[STAG SEQ ID NO:20)]-x(2)-T-x-[DNG]-

10 [1] Martin J.F. Trends Biotechnol. 5:306-308(1987).

[2] Chen G., Shiffman D., Mevarech M., Aharonowitz Y. Trends Biotechnol. 8:105-111(1990).

[3] Samson S.M., Dotzla J.E., Slisz M.L., Becker G.W., van Frank R.M., Veal L.E., Yeh W.K., Miller J.R., Queener S.W., Ingolia T.D. Bio/Technology 5:1207-1214(1987).

15 [4] Kovacevic S., Weigel B.J., Tobin M.B., Ingolia T.D., Miller J.R. J. Bacteriol. 171:754-760(1989).

200. Fibrillarin signature

20 Fibrillarin [1] is a component of a nucleolar small nuclear ribonucleoprotein(SnRNP) particle thought to participate in the first step of the processing of pre-rRNA. In mammals, fibrillarin is associated with the U3, U8 and U13 small nuclear RNAs [2]. Fibrillarin is an extremely well conserved protein of about 320 amino acid residues. Structurally it consists of three different domains: - An N-terminal domain of about 80 amino acids which is very rich in
25 glycine and contains a number of dimethylated arginine residues (DMA). - A central domain of about 90 residues which resembles that of RNA-binding proteins and contains an octameric sequence similar to the RNP-2 consensus found in such proteins. - A C-terminal alpha-helical domain. A protein evolutionary related to fibrillarin has been found [3] in
archaeobacteria such as *Methanococcus vanniellii* or *voltae*. This protein (*gene flpA*) is
30 involved in pre-rRNA processing. It lacks the Gly/Arg-rich N-terminal domain. As a signature pattern, a region was selected that starts with and encompasses the RNP-2 like octapeptide sequence.

Consensus pattern: [GST]-[LIVMAP SEQ ID NO:253])-V-Y-A-[IV]-E-[FY]-[SA]-x-R-x(2)-R-[DE] -

[1] Aris J.P., Blobel G. Proc. Natl. Acad. Sci. U.S.A. 88:931-935(1991).

5 [2] Bandziulis R.J., Swanson M.S., Dreyfuss G. Genes Dev. 3:431-437(1989).

[3] Agha-Amiri K. J. Bacteriol. 176:2124-2127(1994).

201. Filamin/ABP280 repeat

10 [1] Fucini P, Renner C, Herberhold C, Noegel AA, Holak TA, Nat Struct Biol 1997;4:223-230.

202. Fucosyl transferase

15 This family of Fucosyltransferases are the enzymes transferring fucose from GDP-Fucose to GlcNAc in an alpha1,3 linkage [1].

[1] Breton C, Oriol R, Imberty A; Glycobiology 1998;8:87-94.

20 203. 2Fe-2S ferredoxins, iron-sulfur binding region signature (fer2A)

Ferredoxins [1] are a group of iron-sulfur proteins which mediate electron transfer in a wide variety of metabolic reactions. Ferredoxins can be divided into several subgroups depending upon the physiological nature of the iron sulfur cluster(s) and according to sequence similarities. One of these subgroups are the 2Fe-2S ferredoxins, which are proteins or domains of around one hundred amino acid residues that bind a single 2Fe-2S iron-sulfur cluster. The proteins that are known [2] to belong to this family are listed below. - Ferredoxin from photosynthetic organisms; namely plants and algae where it is located in the chloroplast or cyanelle; and cyanobacteria. - Ferredoxin from archaebacteria of the Halobacterium genus. - Ferredoxin IV (gene pftA) and V (gene fdxD) from Rhodobacter capsulatus. - Ferredoxin in 25 the toluene degradation operon (gene xylT) and naphthalene degradation operon (gene nahT) of Pseudomonas putida. - Hypothetical Escherichia coli protein yfaE. - The N-terminal domain of the bifunctional ferredoxin/ferredoxin reductase electron transfer component of the benzoate 1,2-dioxygenase complex (gene benC) from Acinetobacter calcoaceticus, the 30

toluene 4-monooxygenase complex (gene *tmoF*), the toluate 1,2-dioxygenase system (gene *xylZ*), and the xylene monooxygenase system (gene *xylA*) from *Pseudomonas*. - The N-terminal domain of phenol hydroxylase protein p5 (gene *dmpP*) from *Pseudomonas Putida*. - The N-terminal domain of methane monooxygenase component C (gene *mmoC*) from *Methylococcus capsulatus*. - The C-terminal domain of the vanillate degradation pathway protein *vanB* in a *Pseudomonas* species. - The N-terminal domain of bacterial fumarate reductase iron-sulfur protein (gene *frdB*). - The N-terminal domain of CDP-6-deoxy-3,4-glucoseen reductase (gene *ascD*) from *Yersinia pseudotuberculosis*. - The central domain of eukaryotic succinate dehydrogenase (ubiquinone) iron- sulfur protein. - The N-terminal domain of eukaryotic xanthine dehydrogenase. - The N-terminal domain of eukaryotic aldehyde oxidase. In the 2Fe-2S ferredoxins, four cysteine residues bind the iron-sulfur cluster. Three of these cysteines are clustered together in the same region of the protein. Our signature pattern spans that iron-sulfur binding region.

Consensus pattern: C-{C}-{C}-[GA]-{C}-C-[GAST SEQ ID NO:179)]-{CPDEKRHFYW SEQ ID NO:254)}-C [The three C's are 2Fe-2S ligands]-

[1] Meyer J. Trends Ecol. Evol. 3:222-226(1988).[2] Harayama S., Polissi A., Rekik M. FEBS Lett. 285:85-88(1991).

Adrenodoxin family, iron-sulfur binding region signature (*fer2B*)

Ferredoxins [1] are a group of iron-sulfur proteins which mediate electron transfer in a wide variety of metabolic reactions. Ferredoxins can be divided into several subgroups depending upon the physiological nature of the iron sulfur cluster(s) and according to sequence similarities. One family of ferredoxins groups together the following proteins that all bind a single 2Fe-2S iron-sulfur cluster: - Adrenodoxin (ADX) (adrenal ferredoxin), a vertebrate mitochondrial protein which transfers electrons from adrenodoxin reductase to cytochrome P450_{scc}, which is involved in cholesterol side chain cleavage. - Putidaredoxin (PTX), a *Pseudomonas putida* protein which transfers electrons from putidaredoxin reductase to cytochrome P450-cam, which is involved in the oxidation of camphor. - Terpredoxin [2], a *Pseudomonas* protein which transfers electrons from terpredoxin reductase to cytochrome P450-terp, which is involved in the oxidation of alpha-terpineol. - Rhodocoxin [3], a *Rhodococcus* protein which transfers electrons from rhodocoxin reductase to cytochrome

CYP116 (thcB), which is involved in the degradation of thiocarbamate herbicides. -
Escherichia coli ferredoxin (gene fdx) [4] whose exact function is not yet known. -
Rhodobacter capsulatus ferredoxin VI [5], which may transfer electrons to a yet
uncharacterized oxygenase. - Caulobacter crescentus ferredoxin (gene fdxB) [6]. In these
5 proteins, four cysteine residues bind the iron-sulfur cluster. Three of these cysteines are
clustered together in the same region of the protein. Our signature pattern spans that iron-
sulfur binding region.

Consensus pattern: C-x(2)-[STAQ SEQ ID NO:145]-x-[STAMV SEQ ID NO:255]-C-
10 [STA]-T-C-[HR] [The three C's are 2Fe-2S ligands]-

[1] Meyer J. Trends Ecol. Evol. 3:222-226(1988).

[2] Peterson J.A., Lu J.-Y., Geisselsoder J., Graham-Lorence S., Carmona C., Witney F.,
Lorence M.C. J. Biol. Chem. 267:14193-14203(1992).

15 [3] Nagy I., Schoofs G., Compennolle F., Proost P., Vanderleyden J., De Mot R. J. Bacteriol.
177:676-687(1995).

[4] Ta D.T., Vickery L.E. J. Biol. Chem. 267:11120-11125(1992).

[5] Naud I., Vincon M., Garin J., Gaillard J., Forest E., Jouanneau Y. Eur. J. Biochem.
222:933-939(1994).

20 [6] Amemiya K EMBL/Genbank: X51607.

204. 4Fe-4S ferredoxins, iron-sulfur binding region signature (fer4)

Ferredoxins [1] are a group of iron-sulfur proteins which mediate electron transfer in a wide
25 variety of metabolic reactions. Ferredoxins can be divided into several subgroups depending
upon the physiological nature of the iron-sulfur cluster(s). One of these subgroups are the
4Fe-4S ferredoxins, which are found in bacteria and which are thus often referred as
'bacterial-type' ferredoxins. The structure of these proteins [2] consists of the duplication of a
domain of twenty six amino acid residues; each of these domains contains four cysteine
30 residues that bind to a 4Fe-4S center. A number of proteins have been found [3] that include
one or more 4Fe-4S binding domains similar to those of bacterial-type ferredoxins. These
proteins are listed below (references are only provided for recently determined sequences). -
The iron-sulfur proteins of the succinate dehydrogenase and the fumarate reductase

complexes (EC 1.3.99.1). These enzyme complexes, which are components of the tricarboxylic acid cycle, each contain three subunits: a flavoprotein, an iron-sulfur protein, and a b-type cytochrome. The iron-sulfur proteins contain three different iron-sulfur centers: a 2Fe-2S, a 3Fe-3S and a 4Fe-4S. - *Escherichia coli* anaerobic glycerol-3-phosphate dehydrogenase (EC 1.1.99.5) This enzyme is composed of three subunits: A, B, and C. The C subunit seems to be an iron-sulfur protein with two ferredoxin-like domains in the N-terminal part of the protein. - *Escherichia coli* anaerobic dimethyl sulfoxide reductase. The B subunit of this enzyme (gene *dmsB*) is an iron-sulfur protein with four 4Fe-4S ferredoxin-like domains. - *Escherichia coli* formate hydrogenlyase. Two of the subunits of this oligomeric complex (genes *hycB* and *hycF*) seem to be iron-sulfur proteins that each contain two 4Fe-4S ferredoxin-like domains. - *Methanobacterium formicicum* formate dehydrogenase (EC 1.2.1.2). This enzyme is used by the archaeobacteria to grow on formate. The beta chain of this dimeric enzyme probably binds two 4Fe-4S centers. - *Escherichia coli* formate dehydrogenases N and O (EC 1.2.1.2). The beta chain of these two enzymes (genes *fdnH* and *fdoH*) are iron-sulfur proteins with four 4Fe-4S ferredoxin-like domains. - *Desulfovibrio* periplasmic [Fe] hydrogenase (EC 1.18.99.1). The large chain of this dimeric enzyme binds three 4Fe-4S centers, two of which are located in the ferredoxin-like N-terminal region of the protein. - *Methanobacterium thermoautrophicum* methyl viologen-reducing hydrogenase subunit *mvhB*, which contains six tandemly repeated ferredoxin-like domains and which probably binds twelve 4Fe-4S centers. - *Salmonella typhimurium* anaerobic sulfite reductase (EC 1.8.1.-) [4]. Two of the subunits of this enzyme (genes *asrA* and *asrC*) seem to both bind two 4Fe-4S centers. - A Ferredoxin-like protein (gene *fixX*) from the nitrogen-fixation genes locus of various *Rhizobium* species, and one from the Nif-region of *Azotobacter* species. - The 9 Kd polypeptide of chloroplast photosystem I [5] (gene *psaC*). This protein contains two low potential 4Fe-4S centers, referred as the A and B centers. - The chloroplast *frxB* protein which is predicted to carry two 4Fe-4S centers. - An ferredoxin from a primitive eukaryote, the enteric amoeba *Entamoeba histolytica*. - *Escherichia coli* hypothetical protein *yjjW*, a protein with a N-terminal region belonging to the radical activating enzymes family (see <PDOC00834>) and two potential 4Fe-4S centers. The pattern of cysteine residues in the iron-sulfur region is sufficient to detect this class of 4Fe-4S binding proteins.

Consensus pattern: C-x(2)-C-x(2)-C-x(3)-C-[PEG] [The four C's are 4Fe-4S ligands]-

- [1] Meyer J. Trends Ecol. Evol. 3:222-226(1988).
 [2] Otake E., Ooi T. J. Mol. Evol. 26:257-267(1987).
 [3] Beinert H. FASEB J. 4:2483-2492(1990).
 [4] Huang C.J., Barrett E.L. J. Bacteriol. 173:1544-1553(1991).
 5 [5] Knaff D.B. Trends Biochem. Sci. 13:460-461(1988).

205. NifH/frxC family signatures (fer4_NifH)

Nitrogenase (EC 1.18.6.1) [1] is the enzyme system responsible for biological nitrogen
 10 fixation. Nitrogenase is an oligomeric complex which consists of two components:
 component 1 which contains the active site for the reduction of nitrogen to ammonia and
 component 2 (also called the iron protein). Component 2 is a homodimer of a protein (gene
 nifH) which binds a single 4Fe-4S iron sulfur cluster [2]. In the nitrogen fixation process nifH
 is first reduced by a protein such as ferredoxin; the reduced protein then transfers electrons to
 15 component 1 with the concomitant consumption of ATP. A number of proteins are known to
 be evolutionary related to nifH. These proteins are: - Chloroplast encoded frxC (or chlL)
 protein [3]. FrxC is encoded on the chloroplast genome of some plant species, its exact
 function is not known, but it could act as an electron carrier in the conversion of
 protochlorophyllide to chlorophyllide. - Rhodobacter capsulatus proteins bchL and bchX [4].
 20 These proteins are also likely to play a role in chlorophyll synthesis. There are a number of
 conserved regions in the sequence of these proteins: in the N-terminal section there is an
 ATP-binding site motif 'A' (P-loop) and in the central section there are two conserved
 cysteines which have been shown, in nifH, to be the ligands of the 4Fe-4S cluster. Two
 signatures patterns that correspond to the regions around these cysteines were developed.

25

Consensus pattern: E-x-G-G-P-x(2)-[GA]-x-G-C-[AG]-G [C binds the iron-sulfur center]-
 Consensus pattern: D-x-L-G-D-V-V-C-G-G-F-[AG]-x-P [C binds the iron-sulfur center]-

- [1] Pau R.N. Trends Biochem. Sci. 14:183-186(1989).
 30 [2] Georgiadis M.M., Komiya H., Chakrabarti P., Woo D., Kornuc J.J., Rees D.C. Science
 257:1653-1659(1992).
 [3] Fujita Y., Takahashi Y., Kohchi T., Ozeki H., Ohyama K., Matsubara H. Plant Mol. Biol.
 13:551-561(1989).

[4] Burke D.H., Alberti M., Hearst J.E. J. Bacteriol. 175:2407-2413(1993).

206. Ferritin iron-binding regions signatures

5 Ferritin [1,2] is one of the major non-heme iron storage proteins. It consists of a mineral core of hydrated ferric oxide, and a multi-subunit protein shell which englobes the former and assures its solubility in an aqueous environment. In animals the protein is mainly cytoplasmic and there are generally two or more genes that encodes for closely related subunits (in mammals there are two subunits which are known as H(eavy) and L(ight)). In plants ferritin
10 is found in the chloroplast [3]. There are a number of well conserved region in the sequence of ferritins. Two of these regions to develop signature patterns were selected. The first pattern is located in the central part of the sequence of ferritin and it contains three conserved glutamate which are thought to be involved in the binding of iron. The second pattern is located in the C-terminal section, it corresponds to a region which forms a hydrophilic channel through
15 which small molecules and ions can gain access to the central cavity of the molecule; this pattern also includes conserved acidic residues which are potential metal-binding sites.

Consensus pattern: E-x-[KR]-E-x(2)-E-[KR]-[LF]-[LIVMA SEQ ID NO:30])-x(2)-Q-N-x-R-x-G-R [The 3 E's are potential iron ligands]-

20 Consensus pattern: D-x(2)-[LIVMF SEQ ID NO:2])-[STAC SEQ ID NO:204])-[DH]-F-[LI]-[EN]-x(2)-[FY]-L-x(6)- [LIVM SEQ ID NO:4])-[KN] [The second D and the E are potential iron ligands]-

[1] Crichton R.R., Charleaux-Wauters M. Eur. J. Biochem. 164:485-506(1987).

25 [2] Theil E.C. Annu. Rev. Biochem. 56:289-315(1987).

[3] Ragland M., Briat J.-F., Gagnon J., Laulhere J.-P., Massenet O., Theil E.C. J. Biol. Chem. 265:18339-18344(1990).

30 207. Intermediate filaments signature (filament)

Intermediate filaments (IF) [1,2,3] are proteins which are primordial components of the cytoskeleton and the nuclear envelope. They generally form filamentous structures 8 to 14 nm wide. IF proteins are members of a very large multigene family of proteins which has

been subdivided in five major subgroups: - Type I: Acidic cytokeratins. - Type II: Basic cytokeratins. - Type III: Vimentin, desmin, glial fibrillary acidic protein (GFAP), peripherin, and plastinin. - Type IV: Neurofilaments L, H and M, alpha-internexin and nestin. - Type V: Nuclear lamins A, B1, B2 and C. All IF proteins are structurally similar in that they consist of: a central rod domain comprising some 300 to 350 residues which is arranged in coiled-coiled alpha-helices, with at least two short characteristic interruptions; a N-terminal non-helical domain (head) of variable length; and a C-terminal domain (tail) which is also non-helical, and which shows extreme length variation between different IF proteins. While IF proteins are evolutionary and structurally related, they have limited sequence homologies except in several regions of the rod domain. A conserved region at the C-terminal extremity of the rod domain was used as a sequence pattern for this class of proteins.

Consensus pattern: [IV]-x-[TACI SEQ ID NO:256)]-Y-[RKH]-x-[LM]-L-[DE]-

[1] Quinlan R., Hutchison C., Lane B. Protein Prof. 2:801-952(1995).

[2] Steiner P.M., Roop D.R. Annu. Rev. Biochem. 57:593-625(1988).

[3] Stewart M. Curr. Opin. Cell Biol. 2:91-100(1990).

208. Flavodoxin signature

Flavodoxins [1,E1] are electron-transfer proteins that function in various electron transport systems. Flavodoxins bind one FMN molecule, which serves as a redox-active prosthetic group. Flavodoxins are functionally interchangeable with ferredoxins. They have been isolated from prokaryotes, cyanobacteria, and some eukaryotic algae. The signature pattern for these proteins is derived from a conserved region in their N-terminal section, this region is involved in the binding of the FMN phosphate group.

Consensus pattern: [LIV]-[LIVFY SEQ ID NO:257)]-[FY]-x-[ST]-x(2)-[AGC]-x-T-x(3)-A-x(2)-[LIV]-

[1] Wakabayashi S., Kimura K., Matsubara H., Rogers L.J. Biochem. J. 263:981-984(1989).

209. Growth factor and cytokines receptors family signatures (fn3)

A number of receptors for lymphokines, hematopoietic growth factors and growth hormone-related molecules have been found [1 to 5] to share a common binding domain. Receptors known to belong to this family are: - Cytokine receptor common beta chain. This chain is common to the IL-3, IL-5 and GM-CSF receptors. - Cytokine receptor common gamma chain. This chain is common to the IL-2, IL-4, IL-7 and IL-13 receptors. - Ciliary neurotrophic factor receptor (CNTFR). - Erythropoietin receptor (EPOR). - Granulocyte colony-stimulating factor receptor (G-CSFR). - Granulocyte-macrophage colony-stimulating factor receptor alpha chain (GM-CSFR). - Interleukin-2 receptor beta chain (IL2R-beta). - Interleukin-3 receptor alpha chain (IL3R). - Interleukin-4 receptor alpha chain (IL4R). - Interleukin-5 receptor alpha chain (IL5R). - Interleukin-6 receptor (IL6R). - Interleukin-7 receptor alpha chain (IL7R). - Interleukin-9 receptor (IL9R). - Growth hormone receptor (GRHR). - Prolactin receptor (PRLR). - Thrombopoietin receptor (TPOR). The conserved region constitutes all or part of the extracellular ligand-binding region and is about 200 amino acid residues long. In the N-terminal of this domain there are two pairs of cysteines known, in the growth hormone receptor, to be involved in disulfide bonds. +-----
 -----xxxxxxx-----+ | C C C C Extracellular XXXXXXXX Cytoplasmic | +-
 |-|-----|-|-----xxxxxxx-----+ ||| Transmembrane +-+ +-+
 + Two patterns to detect this family of receptors were used. The first one is derived from the first N-terminal disulfide loop, the second is a tryptophan-rich pattern located at the C-terminal extremity of the extracellular region.

Consensus pattern: C-[LVFYR SEQ ID NO:258)]-x(7,8)-[STIVDN SEQ ID NO:259)]-C-x-W [The two C's are linked by a disulfide bond]-

Consensus pattern: [STGL SEQ ID NO:260)]-x-W-[SG]-x-W-S-

[1] Bazan J.F. Biochem. Biophys. Res. Commun. 164:788-795(1989).

[2] Bazan J.F. Proc. Natl. Acad. Sci. U.S.A. 87:6934-6938(1990).

[3] Cosman D., Lyman S.D., Idzerda R.L., Beckmann M.P., Park L.S., Goodwin R.G.,

March C.J. Trends Biochem. Sci. 15:265-270(1990).

[4] d'Andrea A.D., Fasman G.D., Lodish H.F. Cell 58:1023-1024(1989).

[5] d'Andrea A.D., Fasman G.D., Lodish H.F. Curr. Opin. Cell Biol. 2:648-651(1990).

210. Phosphoribosylglycinamide formyltransferase active site (formyl_transf)

Phosphoribosylglycinamide formyltransferase (EC 2.1.2.2) (GART) [1] catalyzes the third step in de novo purine biosynthesis, the transfer of a formyl group to 5'-

5 phosphoribosylglycinamide. In higher eukaryotes, GART is part of a multifunctional enzyme polypeptide that catalyzes three of the steps of purine biosynthesis. In bacteria, plants and yeast, GART is a monofunctional protein of about 200 amino-acid residues. In the Escherichia coli enzyme, an aspartic acid residue has been shown to be involved in the catalytic mechanism. The region around this active site residue is well conserved in GART
10 from prokaryotic and eukaryotic sources and can be used as a signature pattern. Mammalian formyltetrahydrofolate dehydrogenase (EC 1.5.1.6) [2] is a cytosolic enzyme responsible for the NADP-dependent decarboxylative reduction of 10-formyltetrahydrofolate into tetrahydrofolate. It is a protein of about 900 amino acids consisting of three domains; the N-terminal domain (200 residues) is structurally related to GARTs. Escherichia coli methionyl-
15 tRNA formyltransferase (EC 2.1.2.9) (gene fnt) [3] is the enzyme responsible for modifying the free amino group of the aminoacyl moiety of methionyl-tRNA (fMet). The central part of fnt seems to be evolutionary related to GART's active site region.

Consensus pattern: G-x-[STM]-[IVT]-x-[FYWVQ SEQ ID NO:261)]-[VMAT SEQ ID
20 NO:262)]-x-[DEVM SEQ ID NO:263)]-x-[LIVMY SEQ ID NO:141)]-D-x-G-x(2)-[LIVT SEQ ID NO:165)]-x(6)-[LIVM SEQ ID NO:4)] [D is the active site residue] -

[1] Inglese J., Smith J.M., Benkovic S.J. Biochemistry 29:6678-6687(1990).

[2] Cook R.J., Lloyd R.S., Wagner C. J. Biol. Chem. 266:4965-4973(1991).

25 [3] Guillon J.-M., Mechulam Y., Schmitter J.-M., Blanquet S., Fayat G. J. Bacteriol. 174:4294-4301(1992).

211. G10 protein signatures

30 A Xenopus protein known as G10 [1] has been found to be highly conserved in a wide range of eukaryotic species. The function of G10 is still unknown. G10 is a protein of about 17 to 18 Kd (143 to 157 residues) which is hydrophilic and whose C-terminal half is rich in

235

cysteines and could be involved in metal-binding. As signature patterns, two of these cysteine-rich segments were selected.

Consensus pattern: L-C-C-x-[KR]-C-x(4)-[DE]-x-N-x(4)-C-x-C-R-V-P-

5 Consensus pattern: C-x-H-C-G-C-[KRH]-G-C-[SA]-

[1] McGrew L.L., Dworkin-Rastl E., Dworkin M.B., Richter J.D. Genes Dev. 3:803-815(1989).

10

212. G-protein alpha subunit

G proteins couple receptors of extracellular signals to intracellular signaling pathways. The G protein alpha subunit binds guanyl nucleotide and is a weak GTPase. Number of members: 195

15

[1] Coleman DE, Berghuis AM, Lee E, Linder ME, Gilman AG, Sprang SR, Science 1994;265:1405-1412.

[2] How G proteins work: a continuing story. Coleman DE, Sprang SR, Trends Biochem Sci 1996;21:41-44.

20

213. Glucose-6-phosphate dehydrogenase active site (G6PD)

Glucose-6-phosphate dehydrogenase (EC 1.1.1.49) (G6PD) [1] catalyzes the first step in the pentose pathway, the reduction of glucose-6-phosphate to gluconolactone 6-phosphate. A lysine residue has been identified as are active nucleophile associated with the activity of the enzyme. The sequence around this lysine is totally conserved from bacterial to mammalian G6PD's and can be used as a signature pattern

25

Consensus pattern: D-H-Y-L-G-K-[EQK] [K is the active site residue]-

30

[1] Jeffery J., Persson B., Wood I., Bergman T., Jeffery R., Joernvall H. Eur. J. Biochem. 212:41-49(1993).

214. GATA-type zinc finger domain

The GATA family of transcription factors are proteins that bind to DNA sites with the consensus sequence (A/T)GATA(A/G), found within the regulatory region of a number of genes. Proteins currently known to belong to this family are: - GATA-1 [1] (also known as Eryf1, GF-1 or NF-E1), which binds to the GATA region of globin genes and other genes expressed in erythroid cells. It is a transcriptional activator which probably serves as a general 'switch' factor for erythroid development. - GATA-2 [2], a transcriptional activator which regulates endothelin-1 gene expression in endothelial cells. - GATA-3 [3], a transcriptional activator which binds to the enhancer of the T-cell receptor alpha and delta genes. - GATA-4 [4], a transcriptional activator expressed in endodermally derived tissues and heart. - Drosophila protein pannier (or DGATAa) (gene pnr) which acts as a repressor of the achaete-scute complex (as-c). - Bombyx mori BCFI [5], which regulates the expression of chorion genes. - Caenorhabditis elegans elt-1 and elt-2, transcriptional activators of genes containing the GATA region, including vitellogenin genes [6]. - Ustilago maydis urbs1 [7], a protein involved in the repression of the biosynthesis of siderophores. - Fission yeast protein GAF2. All these transcription factors contain a pair of highly similar 'zinc finger' type domains with the consensus sequence C-x₂-C-x₁₇-C-x₂-C. Some other proteins contain a single zinc finger motif highly related to those of the GATA transcription factors. These proteins are: - Drosophila box A-binding factor (ABF) (also known as protein serpent (gene srp)) which may function as a transcriptional activator protein and may play a key role in the organogenesis of the fat body. - Emericella nidulans areA [8], a transcriptional activator which mediates nitrogen metabolite repression. - Neurospora crassa nit-2 [9], a transcriptional activator which turns on the expression of genes coding for enzymes required for the use of a variety of secondary nitrogen sources, during conditions of nitrogen limitation. - Neurospora crassa white collar proteins 1 and 2 (WC-1 and WC-2), which control expression of light-regulated genes. - Saccharomyces cerevisiae DAL81 (or UGA43), a negative nitrogen regulatory protein. - Saccharomyces cerevisiae GLN3, a positive nitrogen regulatory protein. - Saccharomyces cerevisiae GAT1. - Saccharomyces cerevisiae GZF3.

Consensus pattern: C-x-[DN]-C-x(4,5)-[ST]-x(2)-W-[HR]-[RK]-x(3)-[GN]-x(3,4)-C-N-[AS]-C [The four C's are zinc ligands]

- [1] Trainor C.D., Evans T., Felsenfeld G., Boguski M.S. *Nature* 343:92-96(1990).
- [2] Lee M.E., Temizer D.T., Clifford J.A., Quertermous T. *J. Biol. Chem.* 266:16188-16192(1991).
- [3] Ho I.-C., Vorhees P., Marin N., Oakley B.K., Tsai S.-F., Orkin S.H., Leiden J.M. *EMBO J.* 10:1187-1192(1991).
- [4] Spieth J., Shim Y.H., Lea K., Conrad R., Blumenthal T. *Mol. Cell. Biol.* 11:4651-4659(1991).
- [5] Drevet J.R., Skeiky Y.A., Iatrou K. *J. Biol. Chem.* 269:10660-10667(1994).
- [6] Hawkins M.G., McGhee J.D. *J. Biol. Chem.* 270:14666-14671(1995).
- [7] Voisard C.P.O., Wang J., Xu P., Leong S.A., McEvoy J.L. *Mol. Cell. Biol.* 13:7091-7100(1993).
- [8] Arst H.N. Jr., Kudla B., Martinez-Rossi N.M., Caddick M.X., Sibley S., Davies R.W. *Trends Genet.* 5:291-291(1989).
- [9] Fu Y.-H., Marzluf G.A. *Mol. Cell. Biol.* 10:1056-1065(1990).

215. Glutamine amidotransferases class-I active site (GATase)

A large group of biosynthetic enzymes are able to catalyze the removal of the ammonia group from glutamine and then to transfer this group to a substrate to form a new carbon-nitrogen group. This catalytic activity is known as glutamine amidotransferase (GATase) (EC 2.4.2.-)

[1]. The GATase domain exists either as a separate polypeptidic subunit or as part of a larger polypeptide fused in different ways to a synthase domain. On the basis of sequence

similarities two classes of GATase domains have been identified [2,3]: class-I(also known as trpG-type) and class-II (also known as purF-type). Class-I GATase domains have been found

in the following enzymes: - The second component of anthranilate synthase (AS) (EC 4.1.3.27) [4]. AS catalyzes the biosynthesis of anthranilate from chorismate and glutamine.

AS is generally a dimeric enzyme: the first component can synthesize anthranilate using ammonia rather than glutamine, whereas component II provides the GATase activity. In some bacteria and in fungi the GATase component of AS is part of a multifunctional protein that

also catalyzes other steps of the biosynthesis of tryptophan. - The second component of 4-amino-4-deoxychorismate (ADC) synthase (EC 4.1.3. -), a dimeric prokaryotic enzyme that function in the pathway that catalyzes the biosynthesis of para-aminobenzoate (PABA) from chorismate and glutamine. The second component (gene pabA) provides the GATase activity

[4]. - CTP synthase (EC 6.3.4.2). CTP synthase catalyzes the final reaction in the biosynthesis of pyrimidine, the ATP-dependent formation of CTP from UTP and glutamine. CTP synthase is a single chain enzyme that contains two distinct domains; the GATase domain is in the C-terminal section [2]. - GMP synthase (glutamine-hydrolyzing) (EC 6.3.5.2). GMP synthase catalyzes the ATP-dependent formation of GMP from xanthosine 5'-phosphate and glutamine. GMP synthase is a single chain enzyme that contains two distinct domains; the GATase domain is in the N-terminal section [5]. - Glutamine-dependent carbamoyl-phosphate synthase (EC 6.3.5.5) (GD-CPSase); an enzyme involved in both arginine and pyrimidine biosynthesis and which catalyzes the ATP-dependent formation of carbamoyl phosphate from glutamine and carbon dioxide. In bacteria GD-CPSase is composed of two subunits: the large chain (gene *carB*) provides the CPSase activity, while the small chain (gene *carA*) provides the GATase activity. In yeast the enzyme involved in arginine biosynthesis is also composed of two subunits: CPA1 (GATase), and CPA2 (CPSase). In most eukaryotes, the first three steps of pyrimidine biosynthesis are catalyzed by a large multifunctional enzyme (called URA2 in yeast, rudimentary in *Drosophila*, and CAD in mammals). The GATase domain is located at the N-terminal extremity of this polypeptide [6]. - Phosphoribosylformylglycinamide synthase II (EC 6.3.5.3), an enzyme that catalyzes the fourth step in the de novo biosynthesis of purines. In some species of bacteria, FGAM synthase II is composed of two subunits: a small chain (gene *purQ*) which provides the GATase activity and a large chain (gene *purL*) which provides the aminator activity. - The histidine amidotransferase *hisH*, an enzyme that catalyzes the fifth step in the biosynthesis of histidine in prokaryotes. In the second component of AS a cysteine has been shown [7] to be essential for the amidotransferase activity. The sequence around this residue is well conserved in all the above GATase domains and can be used as a signature pattern for class-I GATase.-

Consensus pattern: [PAS]-[LIVMFYT SEQ ID NO:143)]-[LIVMFY SEQ ID NO:18)]-G-[LIVMFY SEQ ID NO:18)]-C-[LIVMFYN SEQ ID NO:264)]-G-x-[QEH]- x-[LIVMFA SEQ ID NO:81)] [C is the active site residue]-

- [1] Buchanan J.M. Adv. Enzymol. 39:91-183(1973).
- [2] Weng M., Zalkin H. J. Bacteriol. 169:3023-3028(1987).
- [3] Nyunoya H., Lusty C.J. J. Biol. Chem. 259:9790-9798(1984).
- [4] Crawford I.P. Annu. Rev. Microbiol. 43:567-600(1989).

[5] Zalkin H., Argos P., Narayana S.V.L., Tiedeman A.A., Smith J.M. J. Biol. Chem. 260:3350-3354(1985).

[6] Davidson J.N., Chen K.C., Jamison R.S., Musmanno L.A., Kern C.B. BioEssays 15:157-164(1993).

5 [7] Tso J.Y., Hermodson M.A., Zalkin H. J. Biol. Chem. 255:1451-1457(1980).

216. Glutamine amidotransferases class-II active site (GATase_2)

10 A large group of biosynthetic enzymes are able to catalyze the removal of the ammonia group from glutamine and then to transfer this group to a substrate to form a new carbon-nitrogen group. This catalytic activity is known as glutamine amidotransferase (GATase) (EC 2.4.2.-) [1]. The GATase domain exists either as a separate polypeptidic subunit or as part of a larger polypeptide fused in different ways to a synthase domain. On the basis of sequence similarities two classes of GATase domains have been identified [2,3]: class-I(also known as

15 trpG-type) and class-II (also known as purF-type). Class-II GATase domains have been found in the following enzymes: - Amido phosphoribosyltransferase (glutamine phosphoribosylpyrophosphate amidotransferase) (EC 2.4.2.14). An enzyme which catalyzes the first step in purine biosynthesis, the transfer of the ammonia group of glutamine to PRPP to form 5-phosphoribosylamine (gene purF in bacteria, ADE4 in yeast). - Glucosamine--

20 fructose-6-phosphate aminotransferase (EC 2.6.1.16). This enzyme catalyzes a key reaction in amino sugar synthesis, the formation of glucosamine 6-phosphate from fructose 6-phosphate and glutamine (gene glmS in Escherichia coli, nodM in Rhizobium, GFA1 in yeast) - Asparagine synthetase (glutamine-hydrolyzing) (EC 6.3.5.4). This enzyme is responsible for the synthesis of asparagine from aspartate and glutamine. A cysteine is

25 present at the N-terminal extremity of the mature form of all these enzymes. The cysteine has been shown, in amido phosphoribosyltransferase [4] and in asparagine synthetase [5] to be important for the catalytic mechanism.

30 Consensus pattern: <x(0,11)-C-[GS]-[IV]-[LIVMFYW SEQ ID NO:26)]-[AG] [C is the active site residue]-

[1] Buchanan J.M. Adv. Enzymol. 39:91-183(1973).

[2] Weng M., Zalkin H. J. Bacteriol. 169:3023-3028(1987).

[3] Nyunoya H., Lusty C.J. J. Biol. Chem. 259:9790-9798(1984).

[4] van Heeke G., Schuster M. J. Biol. Chem. 264:5503-5509(1989).

[5] Vollmer S.J., Switzer R.L., Hermodson M.A., Bower S.G., Zalkin H. J. Biol. Chem. 258:10582-10585(1983).

5

217. GDP dissociation inhibitor (GDI)

[1] Schalk I, Zeng K, Wu SK, Stura EA, Matteson J, Huang M, Tandon A, Wilson IA, Balch WE, Nature 1996;381:42-48.

10

218. Oxidoreductase family (GFO_IDH_MocA)

This family of enzymes utilise NADP or NAD. This family: is called the GFO/IDH/MOCA family in swiss-prot.

15

[1] Kingston RL, Scopes RK, Baker EN, Structure 1996;4:1413-1428.

219. GHMP kinases putative ATP-binding domain

The following kinases contains, in their N-terminal section, a conserved Gly/Ser-rich region which is probably involved in the binding of ATP [1]. These kinases are listed below. - Galactokinase (EC 2.7.1.6). - Homoserine kinase (EC 2.7.1.39). - Mevalonate kinase (EC 2.7.1.36). - Phosphomevalonate kinase (EC 2.7.4.2). This group of kinases was called 'GHMP' (from the first letter of their substrate)

20

Consensus pattern: [LIVM SEQ ID NO:4)]-[PK]-x-[GSTA SEQ ID NO:19)]-x(0,1)-G-L-[GS]-S-S-[GSA]-[GSTAC SEQ ID NO:99)]-

25

[1] Tsay Y.H., Robinson G.W. Mol. Cell. Biol. 11:620-631(1991).

30

220. Glucose inhibited division protein A family signatures (GIDA)

Bacterial glucose inhibited division protein A (gene gidA) is a protein of 70Kd whose function is not yet known and whose sequence is highly conserved. It is evolutionary related

241

to yeast hypothetical protein YGL236C, *Caenorhabditis elegans* hypothetical protein F52H3.2 and a *Bacillus subtilis* protein called gid (and which is different from *B. subtilis* gidA). Two highly conserved regions were selected as signature patterns. Both regions are located in the central region of the protein.

5

Consensus pattern: [GS]-[PT]-x-Y-C-P-S-[LIVM SEQ ID NO:4)]-E-x-K-[LIVM SEQ ID NO:4)]-x-[KR]-

Consensus pattern: A-G-Q-x-[NT]-G-x(2)-G-Y-x-E-[SAG](3)-[QS]-G-[LIVM SEQ ID NO:4)](2)-A-G-[LIVMT SEQ ID NO:1)]-N-A-

10

221. (GLFV_dehydrog)

Glu / Leu / Phe / Val dehydrogenases active site

15

- Glutamate dehydrogenases (EC 1.4.1.2, EC 1.4.1.3, and EC 1.4.1.4) (GluDH) are enzymes that catalyze the NAD- or NADP-dependent reversible deamination of glutamate into alpha-ketoglutarate [1,2]. GluDH isozymes are generally involved with either ammonia assimilation or glutamate catabolism.

20

- Leucine dehydrogenase (EC 1.4.1.9) (LeuDH) is a NAD-dependent enzyme that catalyzes the reversible deamination of leucine and several other aliphatic amino acids to their keto analogues [3].

- Phenylalanine dehydrogenase (EC 1.4.1.20) (PheDH) is a NAD-dependent enzyme that catalyzes the reversible deamidation of L-phenylalanine into phenylpyruvate [4].

25

- Valine dehydrogenase (EC 1.4.1.8) (ValDH) is a NADP-dependent enzyme that catalyzes the reversible deamidation of L-valine into 3-methyl-2-oxobutanoate [5].

30

These dehydrogenases are structurally and functionally related. A conserved lysine residue located in a glycine-rich region has been implicated in the catalytic mechanism. The conservation of the region around this residue allows the derivation of a signature pattern for such type of enzymes.

Consensus pattern[LIV]-x(2)-G-G-[SAG]-K-x-[GV]-x(3)-[DNST SEQ ID NO:265)]-[PL] [K is the active site residue] Sequences known to belong to this class detected by the pattern ALL.

- 5 Note all known sequences from this family have Pro in the last position of the pattern with the exception of yeast GluDH which as Leu.

[1] Britton K.L., Baker P.J., Rice D.W., Stillman T.J. Eur. J. Biochem. 209:851-859(1992).

[2] Benachenhou-Lahfa N., Forterre P., Labedan B. J. Mol. Evol. 36:335-346(1993).

- 10 [3] Nagata S., Tanizawa K., Esaki N., Sakamoto Y., Ohshima T., Tanaka H., Soda K. Biochemistry 27:9056-9062(1988).

[4] Takada H., Yoshimura T., Ohshima T., Esaki N., Soda K. J. Biochem. 109:371-376(1991).

[5] Hutchinson C.R., Tang L. J. Bacteriol. 175:4176-4185(1993).

15

222. GMC oxidoreductases signatures

The following FAD flavoproteins oxidoreductases have been found [1,2] to be evolutionary related. These enzymes, which are called 'GMC oxidoreductases', are listed below. - Glucose

- 20 oxidase (EC 1.1.3.4) (GOX) from *Aspergillus niger*. Reaction catalyzed: glucose + oxygen -> delta-gluconolactone + hydrogen peroxide. - Methanol oxidase (EC 1.1.3.13) (MOX) from fungi. Reaction catalyzed: methanol + oxygen -> acetaldehyde + hydrogen peroxide. -

Choline dehydrogenase (EC 1.1.99.1) (CHD) from bacteria. Reaction catalyzed: choline + unknown acceptor -> betaine acetaldehyde + reduced acceptor. - Glucose dehydrogenase

- 25 (GLD) (EC 1.1.99.10) from *Drosophila*. Reaction catalyzed: glucose + unknown acceptor -> delta-gluconolactone + reduced acceptor. - Cholesterol oxidase (CHOD) (EC 1.1.3.6) from

Brevibacterium sterolicum and *Streptomyces* strain SA-COO. Reaction catalyzed: cholesterol + oxygen -> cholest-4-en-3-one + hydrogen peroxide. - AlkJ [3], an alcohol dehydrogenase

- 30 aldehydes. This family also includes a lyase: - (R)-mandelonitrile lyase (EC 4.1.2.10)

(hydroxynitrile lyase) from plants [4], an enzyme involved in cyanogenesis, the release of hydrogen cyanide from injured tissues. These enzymes are proteins of size ranging from 556 (CHD) to 664 (MOX) amino acid residues which share a number of regions of sequence

similarities. One of these regions, located in the N-terminal section, corresponds to the FAD ADP-binding domain. The function of the other conserved domains is not yet known; two of these domains were selected as signature patterns. The first one is located in the N-terminal section of these enzymes, about 50 residues after the ADP-binding domain, while the second one is located in the central section.

Consensus pattern: [GA]-[RKN]-x-[LIV]-G(2)-[GST](2)-x-[LIVM SEQ ID NO:4)]-N-x(3)-[FYWA SEQ ID NO:138)]- x(2)-[PAG]-x(5)-[DNESH SEQ ID NO:139)]-

Consensus pattern: [GS]-[PSTA SEQ ID NO:140)]-x(2)-[ST]-P-x-[LIVM SEQ ID NO:4)](2)-x(2)-S-G-[LIVM SEQ ID NO:4)]-G-

[1] Cavener D.R. J. Mol. Biol. 223:811-814(1992).

[2] Henikoff S., Henikoff J.G. Genomics 19:97-107(1994).

[3] van Beilen J.B., Eggink G., Enequist H., Bos R., Witholt B. Mol. Microbiol. 6:3121-3136(1992).

[4] Cheng I.P., Poulton J.E. Plant Cell Physiol. 34:1139-1143(1993).

223. (GMP_synt_C)

Glutamine amidotransferases class-I active site

A large group of biosynthetic enzymes are able to catalyze the removal of the ammonia group from glutamine and then to transfer this group to a substrate to form a new carbon-nitrogen group. This catalytic activity is known as glutamine amidotransferase (GATase) (EC 2.4.2.-) [1]. The GATase domain exists either as a separate polypeptidic subunit or as part of a larger polypeptide fused in different ways to a synthase domain. On the basis of sequence similarities two classes of GATase domains have been identified [2,3]: class-I (also known as trpG-type) and class-II (also known as purF-type). Class-I GATase domains have been found in the following enzymes:

- The second component of anthranilate synthase (AS) (EC 4.1.3.27) [4]. AS catalyzes the biosynthesis of anthranilate from chorismate and glutamine. AS is generally a dimeric enzyme: the first component can synthesize anthranilate using ammonia rather than

glutamine, whereas component II provides the GATase activity. In some bacteria and in fungi the GATase component of AS is part of a multifunctional protein that also catalyzes other steps of the biosynthesis of tryptophan.

- The second component of 4-amino-4-deoxychorismate (ADC) synthase (EC 4.1.3. -), a dimeric prokaryotic enzyme that function in the pathway that catalyzes the biosynthesis of para-aminobenzoate (PABA) from chorismate and glutamine. The second component (gene pabA) provides the GATase activity [4].

- CTP synthase (EC 6.3.4.2). CTP synthase catalyzes the final reaction in the biosynthesis of pyrimidine, the ATP-dependent formation of CTP from UTP and glutamine. CTP synthase is a single chain enzyme that contains two distinct domains; the GATase domain is in the C-terminal section [2].

- GMP synthase (glutamine-hydrolyzing) (EC 6.3.5.2). GMP synthase catalyzes the ATP-dependent formation of GMP from xanthosine 5'-phosphate and glutamine. GMP synthase is a single chain enzyme that contains two distinct domains; the GATase domain is in the N-terminal section [5].

- Glutamine-dependent carbamoyl-phosphate synthase (EC 6.3.5.5) (GD-CPSase); an enzyme involved in both arginine and pyrimidine biosynthesis and which catalyzes the ATP-dependent formation of carbamoyl phosphate from glutamine and carbon dioxide. In bacteria GD-CPSase is composed of two subunits: the large chain (gene carB) provides the CPSase activity, while the small chain (gene carA) provides the GATase activity. In yeast the enzyme involved in arginine biosynthesis is also composed of two subunits: CPA1 (GATase), and CPA2 (CPSase). In most eukaryotes, the first three steps of pyrimidine biosynthesis are catalyzed by a large multifunctional enzyme (called URA2 in yeast, rudimentary in *Drosophila*, and CAD in mammals). The GATase domain is located at the N-terminal extremity of this polyprotein [6].

- Phosphoribosylformylglycinamide synthase II (EC 6.3.5.3), an enzyme that catalyzes the fourth step in the de novo biosynthesis of purines. In some species of bacteria, FGAM synthase II is composed of two subunits: a small chain (gene purQ) which provides the GATase activity and a large chain (gene purL) which provides the aminator activity.

- The histidine amidotransferase hisH, an enzyme that catalyzes the fifth step in the biosynthesis of histidine in prokaryotes.

In the second component of AS a cysteine has been shown [7] to be essential for the amidotransferase activity. The sequence around this residue is well conserved in all the above GATase domains and can be used as a signature pattern for class-I GATase.

5 Consensus pattern[PAS]-[LIVMFYT SEQ ID NO:143)]-[LIVMFY SEQ ID NO:18)]-G-[LIVMFY SEQ ID NO:18)]-C-[LIVMFYN SEQ ID NO:264)]-G-x-[QEH]- x-[LIVMFA SEQ ID NO:81)] [C is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for 6 sequences.

10 Note: in the first position of the pattern Pro is found in all cases except in the slime mold GD-CPSase where it is replaced by Ala.

[1] Buchanan J.M. Adv. Enzymol. 39:91-183(1973).

[2] Weng M., Zalkin H. J. Bacteriol. 169:3023-3028(1987).

15 [3] Nyunoya H., Lusty C.J. J. Biol. Chem. 259:9790-9798(1984).

[4] Crawford I.P. Annu. Rev. Microbiol. 43:567-600(1989).

[5] Zalkin H., Argos P., Narayana S.V.L., Tiedeman A.A., Smith J.M. J. Biol. Chem. 260:3350-3354(1985).

[6] Davidson J.N., Chen K.C., Jamison R.S., Musmanno L.A., Kern C.B. BioEssays 15:157-
20 164(1993).

[7] Tso J.Y., Hermodson M.A., Zalkin H. J. Biol. Chem. 255:1451-1457(1980).

224. Glutathione peroxidases signatures (GSHPx)

25 Glutathione peroxidase (EC 1.11.1.9) (GSHPx) [1,2] is an enzyme that catalyzes the reduction of hydroxyperoxides by glutathione. Its main function is to protect against the damaging effect of endogenously formed hydroxyperoxides. In higher vertebrates at least four forms of GSHPx are known to exist: a ubiquitous cytosolic form (GSHPx-1), a gastrointestinal cytosolic for (GSHPx-GI) [3], a plasma secreted form (GSHPx-P) [4], and a
30 epididymal secretory form (GSHPx-EP). In addition to these characterized forms, the sequence of a protein of unknown function [5] has been shown to be evolutionary related to those of GSHPx's. In filarial nematode parasites such as *Brugia pahangi* the major soluble cuticular protein, known as gp29, is a secreted GSHPx which could provide a mechanism of

resistance to the immune reaction of the mammalian host by neutralizing the products of the oxidative burst of leukocytes [6]. *Escherichia coli* protein btuE, a periplasmic protein involved in the transport of vitamin B12, is also evolutionary related to GSHPx's; the significance of this relationship is not yet clear. Selenium, in the form of selenocysteine [7] is part of the catalytic site of GSHPx. The sequence around the selenocysteine residue is moderately well conserved in GSHPx's and the related proteins and can be used as a signature pattern. As a second signature for this family of proteins a highly conserved octapeptide located in the central section of these proteins was selected.

Consensus pattern: [GN]-[RKHNFC SEQ ID NO:266)]-x-[LIVMFC SEQ ID NO:90)]-[LIVMF SEQ ID NO:2)](2)-x-N-[VT]-x-[STC]-x-C- [GA]-x-T [C is the active site selenocysteine residue]

Consensus pattern: [LIV]-[AGD]-F-P-[CS]-[NG]-Q-

[1] Mannervik B. Meth. Enzymol. 113:490-495(1985).

[2] Mullenbach G.T., Tabrizi A., Irvine B.D., Bell G.I., Tainer J.A., Hallewell R.A. Protein Eng. 2:239-246(1988).

[3] Chu F.F., Doroshov J.H., Esworthy R.S. J. Biol. Chem. 268:2571-2576(1993).

[4] Takahashi K., Akasaka M., Yamamoto Y., Kobayashi C., Mizoguchi J., Koyama J. J. Biochem. 108:145-148(1990).

[5] Dunn D.K., Howells D.D., Richardson J., Goldfarb P.S. Nucleic Acids Res. 17:6390-6390(1989).

[6] Cookson E., Blaxter M.L., Selkirk M.E. Proc. Natl. Acad. Sci. U.S.A. 89:5837-5841(1992).

[7] Stadtman T.C. Annu. Rev. Biochem. 59:111-127(1990).

225. (GST)

Glutathione S-transferases

Function: conjugation of reduced glutathione to a variety of targets. Also included in the alignment, but are not GSTs S-crystallins from squid. Similarity to GST was previously noted. Eukaryotic elongation factors 1-gamma. Not known to have GST activity; similarity

247

not previously recognized. Supported by HMM and manual alignment inspection. HSP26 family of stress-related proteins. including auxin-regulated proteins in plants and stringent starvation proteins in *E. coli*. Not known to have GST activity. Similarity not previously recognized. Supported by HMM and manual alignment inspection. Alignment spans entire protein.

226. GTP1/OBG family signature

A widespread family of GTP-binding proteins has been recently characterized [1,2]. This family currently includes: - Mouse and *Xenopus* protein DRG. - Human protein DRG2. - *Drosophila* protein 128up. - Fission yeast protein gtp1. - A *Halobacterium cutirubrum* hypothetical protein in a ribosomal protein gene cluster. - *Bacillus subtilis* protein obg. Obg has been experimentally shown to bind GTP. - *Escherichia coli* hypothetical protein yhbZ. - *Haemophilus influenzae* hypothetical protein HI0877. - *Mycoplasma genitalium* hypothetical protein MG384. - Yeast hypothetical protein YAL036c (FUN11). - Yeast hypothetical protein YGR173w. - *Caenorhabditis elegans* hypothetical protein C02F5.3. The function of the proteins that belong to this family is not yet known. They are polypeptides of about 40 to 48 Kd which contain the five small sequence elements characteristic of GTP-binding proteins [3]. As a signature pattern the region that correspond to the ATP/GTP B motif (also called G-3 in GTP-binding proteins) was selected.

Consensus pattern: D-[LIVM SEQ ID NO:4]-P-G-[LIVM SEQ ID NO:4](2)-[DEY]-[GN]-A-x(2)-G-x-G -

[1] Sazuka T., Tomooka Y., Ikawa Y., Noda M., Kumar S. *Biochem. Biophys. Res. Commun.* 189:363-370(1992).

[2] Hudson J.D., Young P.G. *Gene* 125:191-193(1993).

[3] Bourne H.R., Sanders D.A., McCormick F. *Nature* 349:117-127(1991).

227. (GTP_EFTU1)

ATP/GTP-binding site motif A (P-loop)

From sequence comparisons and crystallographic data analysis it has been shown [1,2,3,4,5,6] that an appreciable proportion of proteins that bind ATP or GTP share a number of more or less conserved sequence motifs. The best conserved of these motifs is a glycine-rich region, which typically forms a flexible loop between a beta-strand and an alpha-helix.

5 This loop interacts with one of the phosphate groups of the nucleotide. This sequence motif is generally referred to as the 'A' consensus sequence [1] or the 'P-loop' [5]. There are numerous ATP- or GTP-binding proteins in which the P-loop is found. Listed below are a number of protein families for which the relevance of the presence of such motif has been noted: - ATP synthase alpha and beta subunits (see <PDOC00137>). - Myosin heavy chains. - Kinesin heavy chains and kinesin-like proteins (see <PDOC00343>). - Dynamins and dynamin-like proteins (see <PDOC00362>). - Guanylate kinase (see <PDOC00670>). - Thymidine kinase (see <PDOC00524>). - Thymidylate kinase (see <PDOC01034>). - Shikimate kinase (see <PDOC00868>). - Nitrogenase iron protein family (nifH/frxC) (see <PDOC00580>). - ATP-binding proteins involved in 'active transport' (ABC transporters) [7] (see <PDOC00185>). - DNA and RNA helicases [8,9,10]. - GTP-binding elongation factors (EF-Tu, EF-1alpha, EF-G, EF-2, etc.). - Ras family of GTP-binding proteins (Ras, Rho, Rab, Ral, Ypt1, SEC4, etc.). - Nuclear protein ran (see <PDOC00859>). - ADP-ribosylation factors family (see <PDOC00781>). - Bacterial dnaA protein (see <PDOC00771>). - Bacterial recA protein (see <PDOC00131>). - Bacterial recF protein (see <PDOC00539>). - Guanine nucleotide-binding proteins alpha subunits (Gi, Gs, Gt, G0, etc.). - DNA mismatch repair proteins mutS family (See <PDOC00388>). - Bacterial type II secretion system protein E (see <PDOC00567>). Not all ATP- or GTP-binding proteins are picked-up by this motif. A number of proteins escape detection because the structure of their ATP-binding site is completely different from that of the P-loop. Examples of such proteins are the E1-E2 ATPases or the glycolytic kinases. In other ATP- or GTP-binding proteins the flexible loop exists in a slightly different form; this is the case for tubulins or protein kinases. A special mention must be reserved for adenylate kinase, in which there is a single deviation from the P-loop pattern: in the last position Gly is found instead of Ser or Thr.

30 -Consensus pattern: [AG]-x(4)-G-K-[ST]-

[1] Walker J.E., Saraste M., Runswick M.J., Gay N.J. EMBO J. 1:945-951(1982).

[2] Moller W., Amons R. FEBS Lett. 186:1-7(1985).

- [3] Fry D.C., Kuby S.A., Mildvan A.S. Proc. Natl. Acad. Sci. U.S.A. 83:907-911(1986).
 [4] Dever T.E., Glynias M.J., Merrick W.C. Proc. Natl. Acad. Sci. U.S.A. 84:1814-1818(1987).
 [5] Saraste M., Sibbald P.R., Wittinghofer A. Trends Biochem. Sci. 15:430-434(1990).
 5 [6] Koonin E.V. J. Mol. Biol. 229:1165-1174(1993).
 [7] Higgins C.F., Hyde S.C., Mimmack M.M., Gileadi U., Gill D.R., Gallagher M.P. J. Bioenerg. Biomembr. 22:571-592(1990).
 [8] Hodgman T.C. Nature 333:22-23(1988) and Nature 333:578-578(1988) (Errata).
 [9] Linder P., Lasko P., Ashburner M., Leroy P., Nielsen P.J., Nishi K., Schnier J., Slonimski
 10 P.P. Nature 337:121-122(1989).
 [10] Gorbalenya A.E., Koonin E.V., Donchenko A.P., Blinov V.M. Nucleic Acids Res. 17:4713-4730(1989).

GTP-binding elongation factors signature (GTP_EFTU2)

15 Elongation factors [1,2] are proteins catalyzing the elongation of peptide chains in protein biosynthesis. In both prokaryotes and eukaryotes, there are three distinct types of elongation factors, as described in the following table: -----

-----	Eukaryotes	Prokaryotes	Function	-----
-----	EF-1alpha	EF-Tu	Binds GTP and an aminoacyl-tRNA; delivers the latter to the A site of ribosomes.	-----
20	EF-1beta	EF-Ts	Interacts with EF-1a/EF-Tu to displace GDP and thus allows the regeneration of GTP-EF-1a.	-----
-----	EF-2	EF-G	Binds GTP and peptidyl-tRNA and translocates the latter from the A site to the P site.	-----

-----The GTP-binding elongation factor family also includes the following proteins: - Eukaryotic peptide chain release factor GTP-binding subunits [3]. These proteins
 25 interact with release factors that bind to ribosomes that have encountered a stop codon at their decoding site and help them to induce release of the nascent polypeptide. The yeast protein was known as SUP2 (and also as SUP35, SUP12 or GST1) and the human homolog as GST1-Hs. - Prokaryotic peptide chain release factor 3 (RF-3) (gene prfC). RF-3 is a class-II RF, a GTP-binding protein that interacts with class I RFs (see <PDOC00607>) and enhance
 30 their activity [4]. - Prokaryotic GTP-binding protein lepA and its homolog in yeast (gene GUF1) and in Caenorhabditis elegans (ZK1236.1). - Yeast HBS1 [5]. - Rat statin S1 [6], a protein of unknown function which is highly similar to EF-1alpha. - Prokaryotic selenocysteine-specific elongation factor selB [7], which seems to replace EF-Tu for the

insertion of selenocysteine directed by the UGA codon. - The tetracycline resistance proteins tetM/tetO [8,9] from various bacteria such as *Campylobacter jejuni*, *Enterococcus faecalis*, *Streptococcus mutans* and *Ureaplasma urealyticum*. Tetracycline binds to the prokaryotic ribosomal 30S subunit and inhibits binding of aminoacyl-tRNAs. These proteins abolish the inhibitory effect of tetracycline on protein synthesis. - Rhizobium nodulation protein nodQ [10]. - *Escherichia coli* hypothetical protein yihK [11]. In EF-1- α , a specific region has been shown [12] to be involved in a conformational change mediated by the hydrolysis of GTP to GDP. This region is conserved in both EF-1 α /EF-Tu as well as EF-2/EF-G and thus seems typical for GTP-dependent proteins which bind non-initiator tRNAs to the ribosome. The pattern developed for this family of proteins include that conserved region.

Consensus pattern: D-[KRSTGANQFYW SEQ ID NO:267)]-x(3)-E-[KRAQ SEQ ID NO:268)]-x-[RKQD SEQ ID NO:269)]-[GC]-[IVMK SEQ ID NO:270)]-[ST]-[IV]-x(2)-[GSTACKRNQ SEQ ID NO:271)]-

[1] Concise Encyclopedia Biochemistry, Second Edition, Walter de Gruyter, Berlin New-York (1988).

[2] Moldave K. Annu. Rev. Biochem. 54:1109-1149(1985).

[3] Stansfield I., Jones K.M., Kushnirov V.V., Dagkesamanskaya A.R., Poznyakovski A.I., Paushkin S.V., Nierras C.R., Cox B.S., Ter-Avanesyan M.D., Tuite M.F. EMBO J. 14:4365-4373(1995).

[4] Grentzmann G., Brechemier-Baey D., Heurgue-Hamard V., Buckingham R.H. J. Biol. Chem. 270:10595-10600(1995).

[5] Nelson R.J., Ziegelhoffer T., Nicolet C., Werner-Washburne M., Craig E.A. Cell 71:97-105(1992).

[6] Ann D.K., Moutsatsos I.K., Nakamura T., Lin H.H., Mao P.-L., Lee M.-J., Chin S., Liem R.K.H., Wang E. J. Biol. Chem. 266:10429-10437(1991).

[7] Forchhammer K., Leinfelder W., Bock A. Nature 342:453-456(1989).

[8] Manavathu E.K., Hiratsuka K., Taylor D.E. Gene 62:17-26(1988).

[9] Leblanc D.J., Lee L.N., Titmas B.M., Smith C.J., Tenover F.C. J. Bacteriol. 170:3618-3626(1988).

[10] Cervantes E., Sharma S.B., Maillet F., Vasse J., Truchet G., Rosenberg C. Mol. Microbiol. 3:745-755(1989).

[11] Plunkett G. III, Burland V.D., Daniels D.L., Blattner F.R. *Nucleic Acids Res.* 21:3391-3398(1993).

[12] Moller W., Schipper A., Amons R. *Biochimie* 69:983-989(1987).

5

228. GTP cyclohydrolase II.

GTP cyclohydrolase II catalyses the first committed step in the biosynthesis of riboflavin.

[1] Richter G, Ritz H, Katzenmeier G, Volk R, Kohnle A, Lottspeich F, Allendorf D, Bacher A, *J Bacteriol* 1993;175:4045-4051.

10

229. Galactose-1-phosphate uridyl transferase signatures (GalP_UDP_transf)

Galactose-1-phosphate uridyl transferase (EC 2.7.7.10) (galT) catalyzes the transfer of an uridyldiphosphate group on galactose (or glucose) 1-phosphate. During the reaction, the uridyl moiety links to a histidine residue. In the *Escherichia coli* enzyme, it has been shown [1] that two histidine residues separated by a single proline residue are essential for enzyme activity. On the basis of sequence similarities, two apparently unrelated families seem to exist. Class-I enzymes are found in eukaryotes as well as some bacteria such as *Escherichia coli* or *Streptomyces lividans*, while class-II enzymes have been found so far only in bacteria such as *Bacillus subtilis* or *Lactobacillus helveticus* [2]. Signature patterns for both families were developed. For class-I enzymes the signature is based on the active site residues. For class-II enzymes a region which also includes two conserved histidines was chosen.

15

20

25

Consensus pattern: F-E-N-[RK]-G-x(3)-G-x(4)-H-P-H-x-Q [The two H's are the active site residues]-

Consensus pattern: D-L-P-I-V-G-G-[ST]-[LIVM SEQ ID NO:4])(2)-[SA]-H-[DEN]-H-[FY]-Q-G-G -

Note: class-I enzymes are structurally related to the HIT family of proteins (see

30

<[PDOC00694](#)

[1] Reichardt J.K.V., Berg P. *Nucleic Acids Res.* 16:9017-9026(1988).

[2] Mollet B., Pilloud N. *J. Bacteriol.* 173:4464-4473(1991).

230. Gamma-thionins family signature

The following small plant proteins are evolutionary related:

- 5 - Gamma-thionins from wheat endosperm (gamma-purothionins) and barley (gamma- hordothionins) which are toxic to animal cells and inhibit protein synthesis in cell free systems [1].
- A flower-specific thionin (FST) from tobacco [2].
- Antifungal proteins (AFP) from the seeds of Brassicaceae species such as radish,
- 10 mustard, turnip and Arabidopsis thaliana [3].
- Inhibitors of insect alpha-amylases from sorghum [4].
- Probable protease inhibitor P322 from potato.
- A germination-related protein from cowpea [5].
- Anther-specific protein SF18 from sunflower [6]. SF18 is a protein that contains a
- 15 gamma-thionin domain at its N-terminus and a proline-rich C- terminal domain.
- Soybean sulfur-rich protein SE60 [7].
- Vicia faba antibacterial peptides fabatin-1 and -2.

In their mature form, these proteins generally consist of about 45 to 50 amino-acid residues. As shown in the following schematic representation, these peptides contain eight

20 conserved cysteines involved in disulfide bonds.

```
+-----+ | +-----+ ||||
xxCxxxxxxxxxCxxxxCxxxCxxxxxxxxCxxxxxCxCxxxC *****|***||
+---|-----+ | +-----+
```

'C': conserved cysteine involved in a disulfide bond.

25 '*': position of the pattern.

Consensus pattern: [KRG]-x-C-x(3)-[SV]-x(2)-[FYWH SEQ ID NO:272])-x-[GF]-x-C-x(5)-C-x(3)-C [The four C's are involved in disulfide bonds]-

30 [1] Bruix M., Jimenez M.A., Santoro J., Gonzalez C., Colilla F.J., Mendez E., Rico M. Biochemistry 32:715-724(1993).

[2] Gu Q., Kawata E.E., Morse M.-J., Wu H.-M., Cheung A.Y. Mol. Gen. Genet. 234:89-96(1992).

[3] Terras F.R.G., Torrekens S., van Leuven F., Osborn R.W., Vanderleyden J., Cammue B.P.A., Broekaert W.F. FEBS Lett. 316:233-240(1993).

[4] Bloch C. Jr., Richardson M. FEBS Lett. 279:101-104(1991).

[5] Ishibashi N., Yamauchi D., Miniamikawa T. Plant Mol. Biol. 15:59-64(1990).

5 [7] Choi Y., Choi Y.D., Lee J.S. Plant Physiol. 101:699-700(1993).

231. Gelsolin. Gelsolin repeat. Number of members: 170

10 [1]Medline: 97433077. The crystal structure of plasma gelsolin: implications for actin severing, capping, and nucleation. Burtinck LD, Koepf EK, Grimes J, Jones EY, Stuart DI, McLaughlin PJ, Robinson RC; Cell 1997;90:661-670.

15 232. Germin family signature

Germins [1] are a family of homopentameric cereal glycoproteins expressed during germination which may play a role in altering the properties of cell walls during germinative growth. It has been shown that wheat and barleygermins act as oxalate oxidases (EC 1.2.3.4), an enzyme that catalyzes the oxidative degradation of oxalate to carbonate and hydrogen
20 peroxide. Germins are highly similar to: - Germin-like proteins from various plants such as rape, violet or white mustard. - Slime mold spherulins 1a and 1b which are proteins that accumulate specifically during spherulation, a process induced by various forms of environmental stress which leads to encystment and dormancy. As a signature pattern the best conserved region was selected: a decapeptide located in the central section of these proteins.

25 Consensus pattern: G-x(4)-H-x-H-P-x-A-x-E-[LIVM SEQ ID NO:4)]-

[1] Lane B.G. FASEB J. 8:294-301(1994).

30 233. (GlutR)

Glutamyl-tRNA reductase signature

Delta-aminolevulinic acid (ALA) is the obligatory precursor for the synthesis of all tetrapyrroles including porphyrin derivatives such as chlorophyll and heme. ALA can be synthesized via two different pathways: the Shemin (or C4) pathway which involves the single step condensation of succinyl-CoA and glycine and which is catalyzed by ALA synthase (EC 2.3.1.37) and via the C5 pathway from the five-carbon skeleton of glutamate. The C5 pathway operates in the chloroplast of plants and algae, in cyanobacteria, in some eubacteria and in archaebacteria.

The initial step in the C5 pathway is carried out by glutamyl-tRNA reductase (GluTR) [1] which catalyzes the NADP-dependent conversion of glutamate-tRNA(Glu) to glutamate-1-semialdehyde (GSA) with the concomitant release of tRNA(Glu) which can then be recharged with glutamate by glutamyl-tRNA synthetase.

GluTR is a protein of about 50 Kd (467 to 550 residues) which contains a few conserved region. The best conserved region is located in positions 99 to 122 in the sequence of known GluTR. This region seems important for the activity of the enzyme. We have developed a signature pattern from that conserved region.

Consensus pattern H-[LIVM SEQ ID NO:4)]-x(2)-[LIVM SEQ ID NO:4)]-[GSTAC SEQ ID NO:99)](3)-[LIVM SEQ ID NO:4)]-[DEQ]-S-[LIVMA SEQ ID NO:30)]-[LIVM SEQ ID NO:4)](2)-[GF]-E-x-[EQR]-[IV]-[LIT]-[STAG SEQ ID NO:20)]-Q-[LIVM SEQ ID NO:4)]-[KR] Sequences known to belong to this class detected by the pattern ALL.

[1] Jahn D., Verkamp E., Soell D. Trends Biochem. Sci. 17:215-218(1992).

234. (Glycoprotease)

Glycoprotease family signature (aka Peptidase_M22)

Glycoprotease (GCP) (EC 3.4.24.57) [1], or o-sialoglycoprotein endopeptidase, is a metalloprotease secreted by *Pasteurella haemolytica* which specifically cleaves O-sialoglycoproteins such as glycophorin A. The sequence of GCP is highly similar to the following uncharacterized proteins:

- *Escherichia coli* hypothetical protein ygjD (ORF-X).
- *Bacillus subtilis* hypothetical protein ydiE.
- *Mycobacterium leprae* hypothetical protein U229E.
- 5 - *Mycobacterium tuberculosis* hypothetical protein MtCY78.10.
- *Synechocystis* strain PCC 6803 hypothetical protein slr0807.
- *Methanococcus jannaschii* hypothetical protein MJ1130.
- *Haloarcula marismortui* hypothetical protein in HSH 3' region.
- Yeast hypothetical protein YKR038c.
- 10 - Yeast hypothetical protein QRI7.

One of the conserved regions contains two conserved histidines. It is possible that this region is involved in coordinating a metal ion such as zinc.

- 15 Consensus pattern[KR]-[GSAT SEQ ID NO:100]-x(4)-[FYWLH SEQ ID NO:273]-
[DQNGK SEQ ID NO:274])-x-P-x-[LIVMFY SEQ ID NO:18])-x(3)-H- x(2)-[AG]-H-
[LIVM SEQ ID NO:4)] Sequences known to belong to this class detected by the pattern ALL.

Note: these proteins belong to family M22 in the classification of peptidases [2,E1].

20 ,

[1] Abdullah K.M., Lo R.Y.C., Mellors A. J. Bacteriol. 173:5597-5603(1991).

[2] Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-228(1995).

25

235. (Glucosamine_iso)

Glucosamine/galactosamine-6-phosphate isomerases signature

- 30 Glucosamine-6-phosphate isomerase (EC 5.3.1.10) (or Glc-6-P deaminase) is the enzyme responsible for the conversion of glucosamine 6-phosphate into fructose6 phosphate [1]. It is the last specific step in the pathway for N-acetylglucosamine (GlcNAC) utilization in bacteria such as *Escherichia coli* (gene nagB) or in fungi such as *Candida albicans* (gene NAG1).Glc-6-P isomerase is evolutionary related to: - A putative *Escherichia coli* galactosamine-6-

256

phosphate isomerase (gene *agaI*) [2]. - *Escherichia coli* hypothetical protein *yieK*. - *Bacillus subtilis* hypothetical protein *ybfT*. As a signature pattern a conserved region located in the central part of these enzymes was selected. This region contains a conserved histidine which has been shown [1], in *nagB*, to be important for the pyranose ring-opening step of the catalytic mechanism

Consensus pattern: [LIVM SEQ ID NO:4)]-x(3)-G-x-[LIT]-x-[LIV]-x-[LIVM SEQ ID NO:4)]-x-G-[LIVM SEQ ID NO:4)]-G-x- [DEN]-G-H-

[1] Oliva G., Fontes M.R.M., Garratt R.C., Altamirano M.M., Calcagno M.L., Horjales E. Structure 3:1323-1332(1995).

[2] Reizer J., Ramseier T.M., Reizer A., Charbit A., Saier M.H. Jr. Microbiology 142:231-250(1996).

236. Pneumovirus attachment glycoprotein G (glycoprotein G)

This family includes attachment proteins from respiratory syncytial virus. Glycoprotein G has not been shown to have any neuraminidase or hemagglutinin activity (Swiss-Prot). The amino terminus is thought to be cytoplasmic, and the carboxyl terminus extracellular. The extracellular region contains four completely conserved cysteine residues.

[1] Johnson PR, Spriggs MK, Olmsted RA, Collins PL, Proc Natl Acad Sci U S A 1987;84:5625-5629.

237. Glycosyl transferases group 1

Mutations in this domain of [Swiss:P37287](#) lead to disease (Paroxysmal Nocturnal haemoglobinuria). Members of this family transfer activated sugars to a variety of substrates, including glycogen, Fructose-6-phosphate and lipopolysaccharides. Members of this family transfer UDP, ADP, GDP or CMP linked sugars. The eukaryotic glycogen synthases may be distant members of this family.

238. Glycosyl transferases (Glycos_transf_2)

Diverse family, transferring sugar from UDP-glucose, UDP-N-acetyl-galactosamine, GDP-mannose or CDP-abequose, to a range of substrates including cellulose, dolichol phosphate and teichoic acids.

5

239. (Glucos_transf_3)

Thymidine and pyrimidine-nucleoside phosphorylases signature

10

Thymidine phosphorylase (EC 2.4.2.4) catalyzes the reversible phosphorolysis of thymidine, deoxyuridine and their analogues to their respective bases and 2-deoxyribose 1-phosphate. This enzyme regulates the availability of thymidine and is therefore essential to nucleic acid metabolism.

15

In *Escherichia coli* (gene *deoA*), the enzyme is a dimer of identical subunits of about 48 Kd [1]. In humans it was first identified as platelet-derived endothelial cell growth factor (PD-ECGF) [E1] before being recognized [2] as thymidine phosphorylase.

20

Bacterial pyrimidine-nucleoside phosphorylase (EC 2.4.2.2) (gene *pdp*) [3] is an enzyme evolutionary and structurally related to thymidine phosphorylase.

A well conserved region of 19 residues located in the N-terminal part of these proteins signature pattern for these enzymes was selected.

25

Consensus pattern S-[GS]-R-[GA]-[LIV]-x(2)-[TA]-[GA]-G-T-x-D-x-[LIV]-E Sequences known to belong to this class detected by the pattern ALL.

30

[1] Walter M.R., Cook W.J., Cole L.B., Short S.A., Koszalka G.W., Krenitsky T.A., Ealick S.E. *J. Biol. Chem.* 265:14016-14022(1990).

[2] Furukawa T., Yoshimura A., Sumizawa T., Haraguchi M., Akiyama S.-I., Fukui K., Yamada Y. *Nature* 356:668-668(1992).

[3] Saxild H.H., Andersen L.N., Hammer K. *J. Bacteriol.* 178:424-434(1996).

240. Glycos_transf_4. Glycosyl transferase. Number of members: 44.

[1] Medline: 95252686. A family of UDP-GlcNAc/MurNAc: polyisoprenol-P GlcNAc/MurNAc-1-P transferases. Lehrman MA; Glycobiology 1994;4:768-771.

5

241. Glycosyl hydrolases family 15. 21 members.

10 242. Glycosyl hydrolases family 16 signature

It has been shown [1] that the following glycosyl hydrolases can be classified into a single family on the basis of sequence similarities: - Bacterial beta-1,3-1,4-glucanases, or lichenases, (EC 3.2.1.73) mainly from *Bacillus* but also from *Clostridium thermocellum* (gene *licB*), *Fibrobacter succinogenes* and *Rhodothermus marinus* (gene *bglA*). - *Bacillus*

15 *circulans* beta-1,3-glucanase A1 (EC 3.2.1.39) (gene *glcA*). - *Lamarinase* (EC 3.2.1.6) from *Clostridium thermocellum* (gene *lam1*). - *Streptomyces coelicolor* agarase (EC 3.2.1.81) (gene *dagA*). - *Alteromonas carrageenovora* kappa-carrageenase (EC 3.2.1.83) (gene *cgkA*). Two closely clustered conserved glutamates have been shown [2] to be involved in the catalytic activity of *Bacillus licheniformis* lichenase. The region was used that contains these
20 residues as a signature pattern.

Consensus pattern: E-[LIV]-D-[LIV]-x(0,1)-E-x(2)-[GQ]-[KRNF SEQ ID NO:275)]-x-[PSTA SEQ ID NO:140)] [The two E's are active site residues]-

25 [1] Henrissat B. Biochem. J. 280:309-316(1991).

[2] Juncosa M., Pons J., Dot T., Querol E., Planas A. J. Biol. Chem. 269:14530-14535(1994).

30 243. Glycosyl hydrolases family 17 signature

It has been shown [1,2] that the following glycosyl hydrolases can be classified into a single family on the basis of sequence similarities: - Glucan endo-1,3-beta-glucosidases (EC 3.2.1.39) (endo-(1->3)-beta- glucanase) from various plants. This enzyme may be involved in

the defense of plants against pathogens through its ability to degrade fungal cell wall polysaccharides. - Glucan 1,3-beta-glucosidase (EC 3.2.1.58) (exo-(1->3)-beta-glucanase) from yeast (gene BGL2). This enzyme may play a role in cell expansion during growth, in cell-cell fusion during mating, and in spore release during sporulation. - Lichenases (EC 3.2.1.73) (endo-(1->3,1->4)-beta-glucanase) from various plants. The best conserved region in the sequence of these enzymes is located in their central section. This region contains a conserved tryptophan residue which could be involved in the interaction with the glucan substrates [2] and it also contains a conserved glutamate which has been shown [3] to act as the nucleophile in the catalytic mechanism. this region was used as a signature pattern.

Consensus pattern: [LIVM SEQ ID NO:4)]-x-[LIVMFYWA SEQ ID NO:41)](3)-[STAG SEQ ID NO:20)]-E-[STA]-G-W-P-[STN]-x-[SAGQ SEQ ID NO:190)] [E is an active site residue]-

[1] Henrissat B. Biochem. J. 280:309-316(1991).

[2] Ori N., Sessa G., Lotan T., Himmelhoch S., Fluhr R. EMBO J. 9:3429-3436(1990).

[3] Varghese J.N., Garrett T.P.J., Colman P.M., Chen L., Hoj P.J., Fincher G.B. Proc. Natl. Acad. Sci. U.S.A. 91:2785-2789(1994).

244. Glyoxalase I signatures

Glyoxalase I (EC 4.4.1.5) (lactoylglutathione lyase) catalyzes the first step of the glyoxal pathway, the transformation of methylglyoxal and glutathione into S-lactoylglutathione which is then converted by glyoxalase II to lactic acid [1]. Glyoxalase I is an ubiquitous enzyme which binds one mole of zinc per subunit. The bacterial and yeast enzymes are monomeric while the mammalian one is homodimeric. The sequence of glyoxalase I is well conserved. In bacteria and mammals, the enzyme is a protein of about 130 to 180 residues while in fungi it is about twice longer. In these organisms the enzyme is built out of the tandem repeat of an homologous domain. Two signature patterns for this family were derived. The first one is located in the N-terminal region while the second one is located in the central section of the protein and contains a conserved histidine that could be implicated in the binding of the zinc atom.

260

Consensus pattern: [HQ]-[IVT]-x-[LIVFY SEQ ID NO:257)]-x-[IV]-x(5)-[STA]-x(2)-F-[YM]-x(2,3)-[LMF]-G-[LMF]-

Consensus pattern: G-[NTKQ SEQ ID NO:276)]-x(0,5)-[GA]-[LVFY SEQ ID NO:277)]-[GH]-H-[IVF]-[CGA]-x-[STAGLE SEQ ID NO:278)]-x(2)-[DNC]-

5

[1] Kim N.-S., Umezawa Y., Ohmura S., Kato S. J. Biol. Chem. 268:11217-11221(1993).

245. (Glypican)

10 Glypicans signature

Glypicans [1,2] are a family of heparan sulfate proteoglycans which are anchored to cell membranes by a glycosylphosphatidylinositol (GPI) linkage. Structurally, these proteins consist of three separate domains:

15

- a) A signal sequence;
- b) An extracellular domain of about 500 residues that contains 12 conserved cysteines probably involved in disulfide bonds and which also contains the sites of attachment of the heparan sulfate glycosaminoglycan side chains;
- 20 c) A C-terminal hydrophobic region which is post-translationally removed after formation of the GPI-anchor.

The proteins known to belong to this family are:

25

- Glypican 1 (GPC1).
- Glypican 2 (GPC2) or cerebroglycan.
- Glypican 3 (GPC3) or OCI-5. In man, defects in GPC3 are the cause of a X-linked genetic disease, Simpson-Galabi-Behmel syndrome (SGBS).
- K-glypican.
- 30 - Glypican 5 (GPC5).
- Drosophila protein dally.

The signature pattern that was developed for glypicans is located in the central section of the extracellular domain and contains five of the conserved cysteines.

Consensus pattern C-x(2)-C-x-G-[LIVM SEQ ID NO:4]-x(4)-P-C-x(2)-[FY]-C-x(2)-[LIVM
5 SEQ ID NO:4]-x(2)-G-C [The C's are probably involved in a disulfide bonds] Sequences
known to belong to this class detected by the pattern ALL, except for dally.

[1] Weksberg R., Squire J.A., Templeton D.M. Nat. Genet. 12:225-227(1996).

[2] Watanabe K., Yamada H., Yamaguchi Y. J. Cell Biol. 130:1207-1218(1995).

10

246. Granins signatures

Granins (chromogranins or secretogranins) [1] are a family of acidic proteins present in the secretory granules of a wide variety of endocrine and neuro-endocrine cells. The exact
15 function(s) of these proteins is not yet known but they seem to be the precursors of
biologically active peptides and/or they may act as helper proteins in the packaging of peptide
hormones and neuropeptides. Three members of this family of proteins show some sequence
similarities: - Chromogranin A (CGA) [2]. CGA is a protein of about 420 residues; it is the
precursor of the peptide pancreastatin which strongly inhibits glucose- induced insulin release
20 from the pancreas. - Secretogranin 1 (chromogranin B). A sulfated protein of about 600
residues. - Secretogranin 2 (chromogranin C). A sulfated protein of about 650 residues. Apart
from their subcellular location and the abundance of acidic residues(Asp and Glu), these
proteins do not share many structural similarities. Only one short region, located in the C-
terminal section, is conserved in all these proteins. Chromogranins A and B share a region of
25 high similarity in their N-terminal section; this region includes two cysteine residues involved
in a disulfide bond

Consensus pattern: [DE]-[SN]-L-[SAN]-x(2)-[DE]-x-E-L-

Consensus pattern: C-[LIVM SEQ ID NO:4](2)-E-[LIVM SEQ ID NO:4](2)-S-[DN]-

30 [STA]-L-x-K-x-S-x(3)-[LIVM SEQ ID NO:4]-[STA]-x-E-C [The two C's are linked by a
disulfide bond]-

[1] Huttner W.B., Gerdes H.-H., Rosa P. Trends Biochem. Sci. 16:27-30(1991).

[2] Simon J.-P., Aunis D. Biochem. J. 262:1-13(1989).

247. grpE protein signature

5 In prokaryotes the grpE protein [1] stimulates, jointly with dnaJ, the ATPase activity of the dnaK chaperone. It seems to accelerate the release of ADP from dnaK thus allowing dnaK to recycle more efficiently. GrpE is a protein of about 22 to 25 Kd. In yeast, an evolutionary related mitochondrial protein(gene GRPE) has been shown [2] to associate with the mitochondrial hsp70protein and to thus play a role in the import of proteins from the
10 cytoplasm. As a signature pattern, the most conserved region of grpE was selected. It is located in the C-terminal section.

Consensus pattern: [FL]-[DN]-[PHEA SEQ ID NO:279)]-x(2)-[HM]-x-A-[LIVMTN SEQ ID NO:280)]-x(16,20)-G-[FY]- x(3)-[DEG]-x(2)-[LIVM SEQ ID NO:4)]-[RI]-x-[SA]-x-V-x-
15 [IV]-

[1] Georgopoulos C., Welch W. Annu. Rev. Cell Biol. 9:601-635(1993).

[2] Bolliger L., Deloche O., Glick B.S., Georgopoulos C., Jenoe P., Kronidou N., Horst M., Morishima N., Schatz G. EMBO J. 13:1998-2006(1994).

20

248. Guanylate kinase signature and profile

Guanylate kinase (EC 2.7.4.8) (GK) [1] catalyzes the ATP-dependent phosphorylation of GMP into GDP. It is essential for recycling GMP and indirectly, cGMP. In prokaryotes (such
25 as Escherichia coli), lower eukaryotes (such as yeast) and in vertebrates, GK is a highly conserved monomeric protein of about 200 amino acids. GK has been shown [2,3,4] to be structurally similar to the following proteins: - Protein A57R (or SalG2R) from various strains of Vaccinia virus. This protein is highly similar to GK, but contains a frameshift mutation in the N-terminal section and could therefore be inactive in that virus. The
30 following proteins are characterized by the presence in their sequence of one or more copies of the DHR domain, a SH3 domain (see <PDOC50002> as well as a C-terminal GK-like domain, these protein are collectively termed MAGUKs (membrane-associated guanylate kinase homologs) [5]: - Drosophila lethal(1)discs large-1 tumor suppressor protein (gene

dlg1). This protein is associated with septate junctions in developing flies and defects in the dlgl gene cause neoplastic overgrowth of the imaginal disks. - Mammalian tight junction protein Zo-1. - A family of mammalian synaptic proteins that seem to interact with the cytoplasmic tail of NMDA receptor subunits. This family currently consist of SAP90/PSD-95, CHAPSYN-110/PSD-93, SAP97/DLG1 and SAP102. - Vertebrate 55 Kd erythrocyte membrane protein (p55). p55 is a palmitoylated, membrane-associated protein of unknown function. - Caenorhabditis elegans protein lin-2, which may play a structural role in the induction of the vulva. - Rat protein CASK. - Human protein DLG2. - Human protein DLG3. There is an ATP-binding site (P-loop) in the N-terminal section of GK. This region is not conserved in the GK-like domain of the above proteins which are therefore unlikely to be kinases. However these proteins retain the residues known, in GK, to be involved in the binding of GMP. As a signature pattern a highly conserved region was selected that contains two arginine and a tyrosine which are involved in GMP-binding

Consensus pattern: T-[ST]-R-x(2)-[KR]-x(2)-[DE]-x(2)-G-x(2)-Y-x-[FY]-[LIVMK SEQ ID NO:281)]-

[1] Stehle T., Schulz G.E. J. Mol. Biol. 224:1127-1141(1992).

[2] Bryant P.J., Woods D.F. Cell 68:621-622(1992).

[3] Goebel M.G. Trends Biochem. Sci. 17:99-99(1992).

[4] Zschocke P.D., Schiltz E., Schulz G.E. Eur. J. Biochem. 213:263-269(1993).

[5] Woods D.F., Bryant P.J. Mech. Dev. 44:85-89(1994).

249. (Glyco_hydro_35)

Glycosyl hydrolases family 35 putative active site

Beta-galactosidases (EC 3.2.1.23) from mammals, fungi, plants and the bacteria

Xanthomonas manihotis are evolutionary related [1,2]. They belong to family 35 in the

classification of glycosyl hydrolases [3,E1].

Mammalian beta-galactosidase is a lysosomal enzyme (gene GLB1) which cleaves the terminal galactose from gangliosides, glycoproteins, and glycosaminoglycans and whose deficiency is the cause of the genetic disease Gm(1) gangliosidosis (Morquio disease type B).

- 5 On of the best conserved regions in these enzymes contains a glutamic acid residue which, on the basis of similarities with other families of glycosyl hydrolases [4], probably acts as the proton donor in the catalytic mechanism. This region was used as a signature pattern.

10 Consensus pattern: G-G-P-[LIVM SEQ ID NO:4](2)-x(2)-Q-x-E-N-E-[FY] [The second E is the putative active site residue] Sequences known to belong to this class detected by the pattern ALL.

- [1] Taron C.H., Benner J.S., Hornstra L.J., Guthrie E.P. Glycobiology 5:603-610(1995).
[2] Carey A.T., Holt K., Picard S., Wilde R., Tucker G.A., Bird C.R., Schuch W., Seymour
15 G.B. Plant Physiol. 108:1099-1107(1995).
[3] Henrissat B., Bairoch A. Biochem. J. 293:781-788(1993).
[4] Henrissat B., Callebaut I., Fabrega S., Lehn P., Mornon J.-P., Davies G. Proc. Natl. Acad. Sci. U.S.A. 92:7090-7094(1995).

20

250. (Glyco_hydro_16)

Glycosyl hydrolases family 16 signature

25

It has been shown [1] that the following glycosyl hydrolases can be classified into a single family on the basis of sequence similarities:

- Bacterial beta-1,3-1,4-glucanases, or lichenases, (EC 3.2.1.73) mainly from Bacillus but also from Clostridium thermocellum (gene licB), Fibrobacter succinogenes and Rhodothermus marinus (gene bglA).
- 30 - Bacillus circulans beta-1,3-glucanase A1 (EC 3.2.1.39) (gene glcA).
- Laminarase (EC 3.2.1.6) from Clostridium thermocellum (gene lam1).
- Streptomyces coelicolor agarase (EC 3.2.1.81) (gene dagA).
- Alteromonas carrageenovora kappa-carrageenase (EC 3.2.1.83) (gene cgkA).

Two closely clustered conserved glutamates have been shown [2] to be involved in the catalytic activity of *Bacillus licheniformis* lichenase. The region that contains these residues as a signature pattern was used.

5

Consensus pattern E-[LIV]-D-[LIV]-x(0,1)-E-x(2)-[GQ]-[KRN SEQ ID NO:275)]-x-[PSTA SEQ ID NO:140)] [The two E's are active site residues]

[1] Henrissat B. *Biochem. J.* 280:309-316(1991).

10 [2] Juncosa M., Pons J., Dot T., Querol E., Planas A. *J. Biol. Chem.* 269:14530-14535(1994).

251. (Glyco_hydro_17)

15 Glycosyl hydrolases family 17 signature
(aka glycosyl_hydro4)

It has been shown [1,2] that the following glycosyl hydrolases can be classified into a single family on the basis of sequence similarities:

20

- Glucan endo-1,3-beta-glucosidases (EC 3.2.1.39) (endo-(1->3)-beta-glucanase) from various plants. This enzyme may be involved in the defense of plants against pathogens through its ability to degrade fungal cell wall polysaccharides.

25 - Glucan 1,3-beta-glucosidase (EC 3.2.1.58) (exo-(1->3)-beta-glucanase) from yeast (gene BGL2). This enzyme may play a role in cell expansion during growth, in cell-cell fusion during mating, and in spore release during sporulation.

- Lichenases (EC 3.2.1.73) (endo-(1->3,1->4)-beta-glucanase) from various plants.

The best conserved region in the sequence of these enzymes is located in their central section.

30

This region contains a conserved tryptophan residue which could be involved in the interaction with the glucan substrates [2] and it also contains a conserved glutamate which has been shown [3] to act as the nucleophile in the catalytic mechanism. This region was used as a signature pattern.

Consensus pattern [LIVM SEQ ID NO:4)]-x-[LIVMFYWA SEQ ID NO:41)](3)-[STAG SEQ ID NO:20)]-E-[STA]-G-W-P-[STN]-x-[SAGQ SEQ ID NO:190)] [E is an active site residue] Sequences known to belong to this class detected by the pattern ALL.

5

[1] Henrissat B. Biochem. J. 280:309-316(1991).

[2] Ori N., Sessa G., Lotan T., Himmelhoch S., Fluhr R. EMBO J. 9:3429-3436(1990).

[3] Varghese J.N., Garrett T.P.J., Colman P.M., Chen L., Hoj P.J., Fincher G.B. Proc. Natl. Acad. Sci. U.S.A. 91:2785-2789(1994).

10

252. (Glyco_hydro_3)

Glycosyl hydrolases family 3 active site

15 It has been shown [1,2] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family:

- Beta glucosidases (EC 3.2.1.21) from the fungi *Aspergillus wentii* (A-3),
Hansenula anomala, *Kluyveromyces fragilis*, *Saccharomycopsis fibuligera*,
20 (BGL1 and BGL2), *Schizophyllum commune* and *Trichoderma reesei* (BGL1).

- Beta glucosidases from the bacteria *Agrobacterium tumefaciens* (Cbg1),
Butyrivibrio fibrisolvens (bglA), *Clostridium thermocellum* (bglB),
Escherichia coli (bglX), *Erwinia chrysanthemi* (bgxA) and *Ruminococcus*
albus.

25 - *Alteromonas* strain O-7 beta-hexosaminidase A (EC 3.2.1.52).

- *Bacillus subtilis* hypothetical protein yzbA.

- *Escherichia coli* hypothetical protein ycfO and HI0959, the corresponding
Haemophilus influenzae protein.

30 One of the conserved regions in these enzymes is centered on a conserved aspartic acid residue which has been shown [3], in *Aspergillus wentii* beta- glucosidase A3, to be implicated in the catalytic mechanism. This region was used as a signature pattern.

267

Consensus pattern[LIVM SEQ ID NO:4)](2)-[KR]-x-[EQK]-x(4)-G-[LIVMFT SEQ ID NO:282)]-[LIVT SEQ ID NO:165)]-[LIVMF SEQ ID NO:2)]- [ST]-D-x(2)-[SGADNI SEQ ID NO:283)] [D is the active site residue] Sequences known to belong to this class detected by the patternALL.

5

[1] Henrissat B. Biochem. J. 280:309-316(1991).

[2] Castle L.A., Smith K.D., Morris R.O. J. Bacteriol. 174:1478-1486(1992).

[3] Bause E., Legler G. Biochim. Biophys. Acta 626:459-465(1980).

10

253. (Glyco_hydro_28)

Polygalacturonase active site (aka PG)

15

Polygalacturonase (EC 3.2.1.15) (PG) (pectinase) [1,2] catalyzes the random hydrolysis of 1,4-alpha-D-galactosiduronic linkages in pectate and other galacturonans. In fruit, polygalacturonase plays an important role in cell wall metabolism during ripening. In plant bacterial pathogens such as *Erwinia carotovora* or *Pseudomonas solanacearum* and fungal pathogens such as *Aspergillus niger*, polygalacturonase is involved in maceration and soft-rotting of plant tissue.

20

Exo-poly-alpha-D-galacturonosidase (EC 3.2.1.82) (exoPG) [3] hydrolyzes peptic acid from the non-reducing end, releasing digalacturonate.

25

Prokaryotic, eukaryotic PG and exoPG share a few regions of sequence similarity. The best conserved of these regions was selected. It is centered on a conserved histidine most probably involved in the catalytic mechanism [4].

30

Consensus pattern[GSDENKRH SEQ ID NO:284)]-x(2)-[VMFC SEQ ID NO:285)]-x(2)-[GS]-H-G-[LIVMAG SEQ ID NO:286)]-x(1,2)- [LIVM SEQ ID NO:4)]-G-S [H is the putative active site residue] Sequences known to belong to this class detected by the patternALL.

Note: these proteins belong to family 28 in the classification of glycosyl hydrolases [5].

[1] Ruttowski E., Labitzke R., Khanh N.Q., Loeffler F., Gottschalk M., Jany K.-D. Biochim. Biophys. Acta 1087:104-106(1990).

[2] Huang J., Schell M.A. J. Bacteriol. 172:3879-3887(1990).

5 [3] He S.Y., Collmer A. J. Bacteriol. 172:4988-4995(1990).

[4] Bussink H.J.D., Buxton F.P., Visser J. Curr. Genet. 19:467-474(1991).

[5] Henrissat B. Biochem. J. 280:309-316(1991).

10 254. (Glyco_hydro_32)

Glycosyl hydrolases family 32 active site

It has been shown [1,2] that the following glycosyl hydrolases can be classified into a single family on the basis of sequence similarities:

15

- Inulinase (EC 3.2.1.7) (or inulase) from the fungi *Kluyveromyces marxianus*.

- Beta-fructofuranosidase (EC 3.2.1.26), commonly known as invertase in fungi and plants and as sucrase in bacteria (gene *sacA* or *scrB*).

20 - Raffinose invertase (EC 3.2.1.26) (gene *rafD*) from *Escherichia coli* plasmid pRSD2.

- Levanase (EC 3.2.1.65) (gene *sacC*) from *Bacillus subtilis*.

One of the conserved regions in these enzymes is located in the N-terminal section and contains an aspartic acid residue which has been shown [3], in yeast invertase to be important for the catalytic mechanism. This region was used as a signature pattern.

25

Consensus pattern H-x(2)-P-x(4)-[LIVM SEQ ID NO:4]-N-D-P-N-G [D is the active site residue]

Sequences known to belong to this class detected by the patternALL.

30

[1] Henrissat B. Biochem. J. 280:309-316(1991).

[2] Gunasekaran P., Karunakaran T., Cami B., Mukundan A.G., Preziosi L., Baratti J. J. Bacteriol. 172:6727-6735(1990).

[3] Reddy V.A., Maley F. J. Biol. Chem. 265:10817-10120(1990).

255. (Glyco_hydro_1)

5 Glycosyl hydrolases family 1 signatures

It has been shown [1 to 4] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family:

- 10 - Beta-glucosidases (EC 3.2.1.21) from various bacteria such as *Agrobacterium* strain ATCC 21400, *Bacillus polymyxa*, and *Caldocellum saccharolyticum*.
- Two plants (clover) beta-glucosidases (EC 3.2.1.21).
- Two different beta-galactosidases (EC 3.2.1.23) from the archaeobacteria *Sulfolobus solfataricus* (genes *bgaS* and *lacS*).
- 15 - 6-phospho-beta-galactosidases (EC 3.2.1.85) from various bacteria such as *Lactobacillus casei*, *Lactococcus lactis*, and *Staphylococcus aureus*.
- 6-phospho-beta-glucosidases (EC 3.2.1.86) from *Escherichia coli* (genes *bglB* and *ascB*) and from *Erwinia chrysanthemi* (gene *arbB*).
- Plants myrosinases (EC 3.2.3.1) (sinigrinase) (thioglucosidase).
- 20 - Mammalian lactase-phlorizin hydrolase (LPH) (EC 3.2.1.108 / EC 3.2.1.62).
LPH, an integral membrane glycoprotein, is the enzyme that splits lactose in the small intestine. LPH is a large protein of about 1900 residues which contains four tandem repeats of a domain of about 450 residues which is evolutionary related to the above glycosyl hydrolases.

25

One of the conserved regions in these enzymes is centered on a conserved glutamic acid residue which has been shown [5], in the beta-glucosidase from *Agrobacterium*, to be directly involved in glycosidic bond cleavage by acting as a nucleophile. This region was used as a signature pattern. As a second signature pattern we selected a conserved region, found in the
30 N-terminal extremity of these enzymes, this region also contains a glutamic acid residue.

Consensus pattern[LIVMFSTC SEQ ID NO:287)]-[LIVFYS SEQ ID NO:288)]-[LIV]-
[LIVMST SEQ ID NO:48)]-E-N-G-[LIVMFAR SEQ ID NO:289)]-[CSAGN SEQ ID

270

NO:290)] [E is the active site residue] Sequences known to belong to this class detected by the patternALL.

Note: this pattern will pick up the last two domains of LPH; the first two domains, which are removed from the LPH precursor by proteolytic processing, have lost the active site glutamate and may therefore be inactive [4].

Consensus patternF-x-[FYWM SEQ ID NO:137)]-[GSTA SEQ ID NO:19)]-x-[GSTA SEQ ID NO:19)]-x-[GSTA SEQ ID NO:19)](2)-[FYNH SEQ ID NO:291)]-[NQ]-x-E-x- [GSTA SEQ ID NO:19)] Sequences known to belong to this class detected by the pattern ALL.

Note: this pattern will pick up the last three domains of LPH.

[1] Henrissat B. Biochem. J. 280:309-316(1991).

[2] Henrissat B. Protein Seq. Data Anal. 4:61-62(1991).

[3] Gonzalez-Candelas L., Ramon D., Polaina J. Gene 95:31-38(1990).

[4] El Hassouni M., Henrissat B., Chippaux M., Barras F. J. Bacteriol. 174:765-777(1992).

[5] Withers S.G., Warren R.A.J., Street I.P., Rupitz K., Kempton J.B., Aebersold R. J. Am. Chem. Soc. 112:5887-5889(1990).

256. Glyco_hydro_20

Glycosyl hydrolase family 20

Previous Pfam IDs: glycosyl_hydr11;

Number of members: 33

257. (Glyco_hydro_9)

Glycosyl hydrolases family 9 active sites signatures

(aka Glycosyl_hydr12)

The microbial degradation of cellulose and xylans requires several types of enzymes such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases

(EC 3.2.1.8) [1,2]. Fungi and bacteria produces a spectrum of cellulolytic enzymes (cellulases) and xylanases which, on the basis of sequence similarities, can be classified into families. One of these families is known as the cellulase family E [3] or as the glycosyl hydrolases family 9 [4,E1]. The enzymes which are currently known to belong to this family are listed below.

- *Butyrivibrio fibrisolvens* cellodextrinase 1 (ced1).
- *Cellulomonas fimi* endoglucanases B (cenB) and C (cenC).
- *Clostridium cellulolyticum* endoglucanase G (celCCG).
- *Clostridium cellulovorans* endoglucanase C (engC).
- *Clostridium stercoararium* endoglucanase Z (avicelase I) (celZ).
- *Clostridium thermocellum* endoglucanases D (celD), F (celF) and I (celI).
- *Fibrobacter succinogenes* endoglucanase A (endA).
- *Pseudomonas fluorescens* endoglucanase A (celA).
- *Streptomyces reticuli* endoglucanase 1 (cel1).
- *Thermomonospora fusca* endoglucanase E-4 (celD).
- *Dictyostelium discoideum* spore germination specific endoglucanase 270-6. This slime mold enzyme may digest the spore cell wall during germination, to release the enclosed amoeba.
- Endoglucanases from plants such as Avocado or French bean. In plants this enzyme may be involved the fruit ripening process.

Two of the most conserved regions in these enzymes are centered on conserved residues which have been shown [5,6], in the endoglucanase D from *Cellulomonas thermocellum*, to be important for the catalytic activity. The first region contains an active site histidine and the second region contains two catalytically important residues: an aspartate and a glutamate. Both regions were used as signature patterns.

Consensus pattern [STV]-x-[LIVMFY SEQ ID NO:18)]-[STV]-x(2)-G-x-[NKR]-x(4)-[PLIVM SEQ ID NO:292)]-H-x-R [H is an active site residue] Sequences known to belong to this class detected by the pattern ALL, except for *Cellulomonas fimi* cenC and *Streptomyces reticuli* cel1.

Consensus pattern [FYW]-x-D-x(4)-[FYW]-x(3)-E-x-[STA]-x(3)-N-[STA] [D and E are active site residues] Sequences known to belong to this class detected by the pattern ALL, except for *Fibrobacter succinogenes* endA whose sequence seems to be incorrect.

5

[1] Beguin P. Annu. Rev. Microbiol. 44:219-248(1990).

[2] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev. 55:303-315(1991).

[3] Henrissat B., Claeysens M., Tomme P., Lemesle L., Mornon J.-P. Gene 81:83-95(1989).

10 [4] Henrissat B. Biochem. J. 280:309-316(1991).

[5] Tomme P., Chauvaux S., Beguin P., Millet J., Aubert J.-P., Claeysens M. J. Biol. Chem. 266:10313-10318(1991).

[6] Tomme P., van Beeumen J., Claeysens M. Biochem. J. 285:319-324(1992).

15

258. Matrix protein (MA), p15 (GAG_ma)

The matrix protein, p15, is encoded by the gag gene. MA is involved in pathogenicity

[1].

[1] : Pozsgay JM, Beilharz MW, Wines BD, Hess AD, Pitha PM, J Virol

20 1993;67:5989-5999.

259. Gag polyprotein, inner coat protein p12 (GAG_P12)

The retroviral p12 is a virion structural protein. p12 is proline rich. The function

25 carried out by p12 in assembly and replication is unknown. p12C is associated with pathogenicity of the virus

[1] Pozsgay JM, Beilharz MW, Wines BD, Hess AD, Pitha PM, J Virol 1993;67:5989-5999.

30 260. Glutamine synthetase signatures (GLN-SYNT)

Glutamine synthetase (EC 6.3.1.2) (GS) [1] plays an essential role in the metabolism of nitrogen by catalyzing the condensation of glutamate and ammonia to form glutamine. There seem to be three different classes of GS [2,3,4]: - Class I enzymes (GSI) are specific to

prokaryotes, and are oligomers of 12 identical subunits. The activity of GSI-type enzyme is controlled by the adenylation of a tyrosine residue. The adenylated enzyme is inactive. - Class II enzymes (GSII) are found in eukaryotes and in bacteria belonging to the Rhizobiaceae, Frankiaceae, and Streptomycetaceae families (these bacteria have also a class-I GS). GSII are octamer of identical subunits. Plants have two or more isozymes of GSII, one of the isozymes is translocated into the chloroplast. - Class III enzymes (GSIII) has, currently, only been found in *Bacteroides fragilis* and in *butyrivibrio fibrisolvens*. It is a hexamer of identical chains. It is much larger (about 700 amino acids) than the GSI (450 to 470 amino acids) or GSII (350 to 420 amino acids) enzymes. While the three classes of GS's are clearly structurally related, the sequence similarities are not so extensive. As signature patterns three conserved regions were selected. The first pattern is based on a conserved tetrapeptide in the N-terminal section of the enzyme, the second one is based on a glycine-rich region which is thought to be involved in ATP-binding. The third pattern is specific to class I glutamine synthetases and includes the tyrosine residue which is reversibly adenylated.

Consensus pattern: [FYWL SEQ ID NO:293]-D-G-S-S-x(6,8)-[DENQSTAK SEQ ID NO:294]-[SA]-[DE]-x(2)-[LIVMFY SEQ ID NO:18)]-

Consensus pattern: K-P-[LIVMFYA SEQ ID NO:98)]-x(3,5)-[NPAT SEQ ID NO:295)]-G-[GSTAN SEQ ID NO:296)]-G-x-H-x(3)-S-

Consensus pattern: K-[LIVM SEQ ID NO:4)]-x(5)-[LIVMA SEQ ID NO:30)]-D-[RK]-[DN]-[LI]-Y [Y is the site of adenylation]-

[1] Eisenberg D., Almassy R.J., Janson C.A., Chapman M.S., Suh S.W., Cascio D., Smith W.W. Cold Spring Harbor Symp. Quant. Biol. 52:483-490(1987).

[2] Kumada Y., Benson D.R., Hillemann D., Hosted T.J., Rochefort D.A., Thompson C.J., Wohlleben W., Tateno Y. Proc. Natl. Acad. Sci. U.S.A. 90:3009-3013(1993).

[3] Shatters R.G., Kahn M.L. J. Mol. Evol. 29:422-428(1989).

[4] Brown J.R., Masuchi Y., Robb F.T., Doolittle W.F. J. Mol. Evol. 38:566-576(1994).

Globins are heme-containing proteins involved in binding and/or transporting oxygen [1].

They belong to a very large and well studied family which is widely distributed in many organisms. The major groups of globins are: - Hemoglobins (Hb) from vertebrates. Hb is the protein responsible for transporting oxygen from the lungs to other tissues. It is a tetramer of

5 two alpha and two beta chains. Most vertebrate species also express specific embryonic or fetal forms of hemoglobin where the alpha or the beta chains are replaced by a chain with higher oxygen affinity, as for the gamma, delta, epsilon and zeta chains in mammals, for example. - Myoglobins (Mg) from vertebrates. Mg is a monomeric protein responsible for

10 oxygen storage in muscles. - Invertebrate globins [2]. A wide variety of globins are found in invertebrates. Molluscs generally have one or two muscle globins which are either

monomeric or dimeric. Insects, such as the midge *Chironomus thummi*, have a large set of extracellular globins. Nematodes and annelids have a variety of intracellular and extracellular globins; some of them are multi- domain polypeptides (from two up to nine-domain globins)

15 and some produce large, disulfide-bonded aggregates. - Leghemoglobins (Lg) from the root nodules of leguminous plants. Lg provides oxygen for bacteroids. - Flavohemoproteins from bacteria (*Escherichia coli hmpA*) and fungi [3]. These proteins consist of two distinct domains: an N-terminal globin domain and a C-terminal FAD-containing reductase domain.

In bacteria such as *Vitreoscilla*, the enzyme-associated globin is a single domain protein. All these globins seem to have evolved from a common ancestor. The profile developed to detect

20 members of the globin family is based on a structural alignment of selected globin sequences [1] Concise Encyclopedia Biochemistry, Second Edition, Walter de Gruyter, Berlin New-

York (1988).[2] Goodman M., Pedwaydon J., Czelusniak J., Suzuki T., Gotoh T., Moens L., Shishikura F., Walz D., Vinogradov S. J. Mol. Evol. 27:236-249(1988).

25 Plant hemoglobins signature (globin2)

Leghemoglobins [1] are hemoproteins present in the root nodules of leguminous plants.

Leghemoglobins are structurally and functionally related to hemoglobin and myoglobin. By providing oxygen to the bacteroids, they are essential for symbiotic nitrogen fixation.

30 Structurally related hemoglobins from the nodules of non-leguminous plants [2,3], and from the roots of non-nodulating plants[4] have been recently sequenced. A signature pattern was developed that picks up the sequence of plants hemoglobins, exclusively.

Consensus pattern: [SN]-P-x-L-x(2)-H-A-x(3)-F-

[1] Powell R., Gannon F. BioEssays 9:117-121(1988).

[2] Kortt A.A., Trinick M.J., Appleby C.A. Eur. J. Biochem. 175:141-149(1988).

[3] Kortt A.A., Inglis A.S., Fleming A.I., Appleby C.A. FEBS Lett. 231:341-346(1988).

5 [4] Bogusz D., Appleby C.A., Landsmann J., Dennis E.S., Trinick M.J., Peacock W.J.
Nature 331:178-180(1988).

262. Fructose-bisphosphate aldolase class-I active site (glycolytic_enz)

10 Fructose-bisphosphate aldolase [1,2] is a glycolytic enzyme that catalyzes the
reversible aldol cleavage or condensation of fructose-1,6-bisphosphate into
dihydroxyacetone-phosphate and glyceraldehyde 3-phosphate. There are two classes of
fructose-bisphosphate aldolases with different catalytic mechanisms. Class-I aldolases [3],
mainly found in higher eukaryotes, are homotetrameric enzymes which form a Schiff-base
15 intermediate between the C-2 carbonyl group of the substrate (dihydroxyacetone
phosphate) and the epsilon-amino group of a lysine residue. In vertebrates, three forms of this
enzyme are found: aldolase A in muscle, aldolase B in liver and aldolase C in brain. The
sequence around the lysine involved in the Schiff-base is highly conserved and can be used as
a signature for this class of enzyme.

20

Consensus pattern: [LIVM SEQ ID NO:4)]-x-[LIVMFYW SEQ ID NO:26)]-E-G-x-[LS]-L-
K-P-[SN] [K is involved in Schiff-base formation]-

[1] Perham R.N. Biochem. Soc. Trans. 18:185-187(1990).

25 [2] Marsh J.J., Lebherz H.G. Trends Biochem. Sci. 17:110-113(1992).

[3] Freemont P.S., Dunbar B., Fothergill-Gilmore L.A. Biochem. J. 249:779-788(1988).

263. Glycosyl hydrolases family 11 active sites signatures

30 The microbial degradation of cellulose and xylans requires several types of enzymes such as
endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases
(EC 3.2.1.8) [1,2]. Fungi and bacteria produces a spectrum of cellulolytic enzymes
(cellulases) and xylanases which, on the basis of sequence similarities, can be classified into

families. One of these families is known as the cellulase family G [3] or as the glycosyl hydrolases family 11 [4,E1]. The enzymes which are currently known to belong to this family are listed below. - *Aspergillus awamori* xylanase C (xynC). - *Bacillus circulans*, *pumilus*, *stearothermophilus* and *subtilis* xylanase (xynA). - *Clostridium acetobutylicum* xylanase (xynB). - *Clostridium stercorarium* xylanase A (xynA). - *Fibrobacter succinogenes* xylanase C (xynC) which consist of two catalytic domains that both belong to family 10. - *Neocallimastix patriciarum* xylanase A (xynA). - *Ruminococcus flavefaciens* bifunctional xylanase XYLA (xynA). This protein consists of three domains: a N-terminal xylanase catalytic domain that belongs to family 11 of glycosyl hydrolases; a central domain composed of short repeats of Gln, Asn an Trp, and a C-terminal xylanase catalytic domain that belongs to family 10 of glycosyl hydrolases. - *Schizophyllum commune* xylanase A. - *Streptomyces lividans* xylanases B (xlnB) and C (xlnC). - *Trichoderma reesei* xylanases I and II. Two of the conserved regions in these enzymes are centered on glutamic acid residues which have both been shown [5], in *Bacillus pumilis* xylanase, to be necessary for catalytic activity. Both regions were used as signature patterns.

Consensus pattern: [PSA]-[LQ]-x-E-Y-Y-[LIVM SEQ ID NO:4])(2)-[DE]-x-[FYWHN SEQ ID NO:297)] [E is an active site residue]-

Consensus pattern: [LIVMF SEQ ID NO:2)]-x(2)-E-[AG]-[YWG]-[QRFGS SEQ ID NO:298)]-[SG]-[STAN SEQ ID NO:250)]-G-x-[SAF] [E is an active site residue]-

[1] Beguin P. Annu. Rev. Microbiol. 44:219-248(1990).

[2] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev. 55:303-315(1991).

[3] Henrissat B., Claeyssens M., Tomme P., Lemesle L., Mornon J.-P. Gene 81:83-95(1989).

[4] Henrissat B. Biochem. J. 280:309-316(1991).

[5] Ko E.P., Akatsuka H., Moriyama H., Shinmyo A., Hata Y., Katsube Y., Urabe I., Okada H. Biochem. J. 288:117-121(1992).

264. Glycosyl hydrolase family 14

This family are beta amylases.

265. Glycosyl hydrolases family 1 signatures

It has been shown [1 to 4] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family: - Beta-glucosidases (EC 3.2.1.21) from

various bacteria such as *Agrobacterium* strain ATCC 21400, *Bacillus polymyxa*, and *Caldocellum saccharolyticum*. - Two plants (clover) beta-glucosidases (EC 3.2.1.21). - Two

different beta-galactosidases (EC 3.2.1.23) from the archaebacteria *Sulfolobus solfataricus* (genes *bgaS* and *lacS*). - 6-phospho-beta-galactosidases (EC 3.2.1.85) from various bacteria

such as *Lactobacillus casei*, *Lactococcus lactis*, and *Staphylococcus aureus*. - 6-phospho-

beta-glucosidases (EC 3.2.1.86) from *Escherichia coli* (genes *bglB* and *ascB*) and from

Erwinia chrysanthemi (gene *arbB*). - Plants myrosinases (EC 3.2.3.1) (sinigrinase)

(thioglucosidase). - Mammalian lactase-phlorizin hydrolase (LPH) (EC 3.2.1.108 / EC

3.2.1.62). LPH, an integral membrane glycoprotein, is the enzyme that splits lactose in the small intestine. LPH is a large protein of about 1900 residues which contains four tandem

repeats of a domain of about 450 residues which is evolutionary related to the above glycosyl

hydrolases. One of the conserved regions in these enzymes is centered on a conserved glutamic acid residue which has been shown [5], in the beta-glucosidase from

Agrobacterium, to be directly involved in glycosidic bond cleavage by acting as a

nucleophile. This region was used as a signature pattern. As a second signature pattern a

conserved region was selected, found in the N-terminal extremity of these enzymes, this region also contains a glutamic acid residue.

Consensus pattern: [LIVMFSTC SEQ ID NO:287]-[LIVFYS SEQ ID NO:288]-[LIV]-

[LIVMST SEQ ID NO:48]-E-N-G-[LIVMFAR SEQ ID NO:289]-[CSAGN SEQ ID

NO:290]] [E is the active site residue]

Note: this pattern will pick up the last two domains of LPH; the first two domains, which are removed from the LPH precursor by proteolytic processing, have lost the active site glutamate and may therefore be inactive [4].

Consensus pattern: F-x-[FYWM SEQ ID NO:137]-[GSTA SEQ ID NO:19]-x-[GSTA SEQ

ID NO:19]-x-[GSTA SEQ ID NO:19]](2)-[FYNH SEQ ID NO:291]-[NQ]-x-E-x- [GSTA SEQ ID NO:19]]-

[2] Henrissat B. Protein Seq. Data Anal. 4:61-62(1991).

[3] Gonzalez-Candelas L., Ramon D., Polaina J. Gene 95:31-38(1990).

[4] El Hassouni M., Henrissat B., Chippaux M., Barras F. J. Bacteriol. 174:765-777(1992).

5 [5] Withers S.G., Warren R.A.J., Street I.P., Rupitz K., Kempton J.B., Aebersold R. J. Am. Chem. Soc. 112:5887-5889(1990).

266. Glycosyl hydrolases family 2 signatures

10 It has been shown [1,2,E1] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family: - Beta-galactosidases (EC 3.2.1.23) from bacteria such as Escherichia coli (genes lacZ and ebgA), Clostridium acetobutylicum, Clostridium thermosulfurogenes, Klebsiella pneumoniae, Lactobacillus delbrueckii, or Streptococcus thermophilus and from the fungi Kluyveromyces lactis. - Beta-glucuronidase
15 (EC 3.2.1.31) from Escherichia coli (gene uidA) and from mammals. One of the conserved regions in these enzymes is centered on a conserved glutamic acid residue which has been shown [3], in Escherichia coli lacZ, to be the general acid/base catalyst in the active site of the enzyme. This region was used as a signature pattern. As a second signature pattern a highly conserved region was selected located some sixty residues upstream from the active site glutamate.

20

Consensus pattern: N-x-[LIVMFYWD SEQ ID NO:299])-R-[STACN SEQ ID NO:300)](2)-H-Y-P-x(4)-[LIVMFYWS SEQ ID NO:301)](2)-x(3)- [DN]-x(2)-G-[LIVMFYW SEQ ID NO:26)](4)-

25 Consensus pattern: [DENQLF SEQ ID NO:302)]-[KRVW SEQ ID NO:303)]-N-[HRY]-[STAPV SEQ ID NO:304)]-[SAC]-[LIVMFS SEQ ID NO:132)](3)-W-[GS]- x(2,3)-N-E [E is the active site residue]-

[1] Henrissat B. Biochem. J. 280:309-316(1991).

30 [2] Schroeder C.J., Robert C., Lenzen G., McKay L.L., Mercenier A. J. Gen. Microbiol. 137:369-380(1991).

[3] Gebler J.C., Aebersold R., Withers S.G. J. Biol. Chem. 267:11126-11130(1992).

267. Glycosyl hydrolases family 3 active site

It has been shown [1,2] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family:

- Beta glucosidases (EC 3.2.1.21) from the fungi *Aspergillus wentii* (A-3),
 5 *Hansenula anomala*, *Kluyveromyces fragilis*, *Saccharomycopsis fibuligera*,
 (BGL1 and BGL2), *Schizophyllum commune* and *Trichoderma reesei* (BGL1).
- Beta glucosidases from the bacteria *Agrobacterium tumefaciens* (Cbg1),
 Butyrivibrio fibrisolvens (bglA), *Clostridium thermocellum* (bglB),
 Escherichia coli (bglX), *Erwinia chrysanthemi* (bgxA) and *Ruminococcus*
 10 *albus*. - *Alteromonas* strain O-7 beta-hexosaminidase A (EC 3.2.1.52).
- *Bacillus subtilis* hypothetical protein yzbA.
- *Escherichia coli* hypothetical protein ycfO and HI0959, the corresponding
 Haemophilus influenzae protein.

One of the conserved regions in these enzymes is centered on a conserved
 15 aspartic acid residue which has been shown [3], in *Aspergillus wentii* beta-
 glucosidase A3, to be implicated in the catalytic mechanism. This
 region was used as a signature pattern.

Consensus pattern: [LIVM SEQ ID NO:4)](2)-[KR]-x-[EQK]-x(4)-G-[LIVMFT SEQ ID
 20 NO:282)]-[LIVT SEQ ID NO:165)]-[LIVMF SEQ ID NO:2)]-[ST]-D-x(2)-[SGADNI SEQ
 ID NO:283)] [D is the active site residue]

- [1] Henrissat B. *Biochem. J.* 280:309-316(1991).
- [2] Castle L.A., Smith K.D., Morris R.O. *J. Bacteriol.* 174:1478-1486(1992).
- 25 [3] Bause E., Legler G. *Biochim. Biophys. Acta* 626:459-465(1980).

268. Glycosyl hydrolases family 8 signature

The microbial degradation of cellulose and xylans requires several types of enzymes such as
 30 endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91)(exoglucanases), or xylanases
 (EC 3.2.1.8) [1,2]. Fungi and bacteria produces a spectrum of cellulolytic enzymes
 (cellulases) and xylanases which, on the basis of sequence similarities, can be classified into
 families. One of these families is known as the cellulase family D [3] or as the glycosyl

hydrolases family 8 [4,E1]. The enzymes which are currently known to belong to this family are listed below. - *Acetobacter xylinum* endonuclease cmcAX. - *Bacillus* strain KSM-330 acidic endonuclease K (Endo-K). - *Cellulomonas josui* endoglucanase 2 (celB). - *Cellulomonas uda* endoglucanase. - *Clostridium cellulolyticum* endoglucanases C (celcCC). - *Clostridium thermocellum* endoglucanases A (celA). - *Erwinia chrysanthemi* minor endoglucanase y (celY). - *Bacillus circulans* beta-glucanase (EC 3.2.1.73). - *Escherichia coli* hypothetical protein yhjM. The most conserved region in these enzymes is a stretch of about 20 residues that contains two conserved aspartate. The first aspartate is thought [5] to act as the nucleophile in the catalytic mechanism. This region was used as a signature pattern.

Consensus pattern: A-[ST]-D-[AG]-D-x(2)-[IM]-A-x-[SA]-[LIVM SEQ ID NO:4]-[LIVMG SEQ ID NO:202])-x-A- x(3)-[FW] [The first D is an active site residue]-

[1] Beguin P. Annu. Rev. Microbiol. 44:219-248(1990).

[2] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev. 55:303-315(1991).

[3] Henrissat B., Claeyssens M., Tomme P., Lemesle L., Mornon J.-P. Gene 81:83-95(1989).

[4] Henrissat B. Biochem. J. 280:309-316(1991).

[5] Alzari P.M., Souchon H., Dominguez R. Structure 4:265-275(1996).

269. Glycosyl hydrolases family 9 active sites signatures

The microbial degradation of cellulose and xylans requires several types of enzymes such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases (EC 3.2.1.8) [1,2]. Fungi and bacteria produce a spectrum of cellulolytic enzymes (cellulases) and xylanases which, on the basis of sequence similarities, can be classified into families.

One of these families is known as the cellulase family E [3] or as the glycosyl hydrolases family 9 [4,E1]. The enzymes which are currently known to belong to this family are listed below. - *Butyrivibrio fibrisolvens* cellodextrinase 1 (ced1). - *Cellulomonas fimi*

endoglucanases B (cenB) and C (cenC). - *Clostridium cellulolyticum* endoglucanase G (celCCG). - *Clostridium cellulovorans* endoglucanase C (engC). - *Clostridium stercoararium* endoglucanase Z (avicelase I) (celZ). - *Clostridium thermocellum* endoglucanases D (celD), F (celF) and I (celI). - *Fibrobacter succinogenes* endoglucanase A (endA). - *Pseudomonas*

fluorescens endoglucanase A (celA). - *Streptomyces reticuli* endoglucanase 1 (cel1). -
 Thermomonospora fusca endoglucanase E-4 (celD). - *Dictyostelium discoideum* spore
 germination specific endoglucanase 270-6. This slime mold enzyme may digest the spore cell
 wall during germination, to release the enclosed amoeba. - Endoglucanases from plants such
 5 as Avocado or French bean. In plants this enzyme may be involved the fruit ripening process.
 Two of the most conserved regions in these enzymes are centered on conserved residues
 which have been shown [5,6], in the endoglucanase D from *Cellulomonas thermocellum*, to
 be important for the catalytic activity. The first region contains an active site histidine and the
 second region contains two catalytically important residues: an aspartate and a glutamate.
 10 Both regions were used as signature patterns.

Consensus pattern: [STV]-x-[LIVMFY SEQ ID NO:18)]-[STV]-x(2)-G-x-[NKR]-x(4)-
 [PLIVM SEQ ID NO:292)]-H-x-R [H is an active site residue]-

Consensus pattern: [FYW]-x-D-x(4)-[FYW]-x(3)-E-x-[STA]-x(3)-N-[STA] [D and E are
 15 active site residues]-

[1] Beguin P. Annu. Rev. Microbiol. 44:219-248(1990).

[2] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev.
 55:303-315(1991).

20 [3] Henrissat B., Claeyssens M., Tomme P., Lemesle L., Mornon J.-P. Gene 81:83-95(1989).

[4] Henrissat B. Biochem. J. 280:309-316(1991).

[5] Tomme P., Chauvaux S., Beguin P., Millet J., Aubert J.-P., Claeyssens M. J. Biol. Chem.
 266:10313-10318(1991).

[6] Tomme P., van Beeumen J., Claeyssens M. Biochem. J. 285:319-324(1992).

25 270. Glyceraldehyde 3-phosphate dehydrogenase active site (gpdh)

Glyceraldehyde 3-phosphate dehydrogenase (EC 1.2.1.12) (GAPDH) [1] is a tetrameric
 NAD-binding enzyme common to both the glycolytic and gluconeogenic pathways. A
 30 cysteine in the middle of the molecule is involved in forming a covalent phosphoglycerol
 thioester intermediate. The sequence around this cysteine is totally conserved in eubacterial
 and eukaryotic GAPDHs and is also present, albeit in a variant form, in the otherwise highly

divergent archaeobacterial GAPDH [2]. Escherichia coli D-erythrose 4-phosphate dehydrogenase (E4PDH) (gene *epd* *orgapB*) is an enzyme highly related to GAPDH [3].

Consensus pattern: [ASV]-S-C-[NT]-T-x(2)-[LIM] [C is the active site residue]-

[1] Harris J.I., Waters M. (In) The Enzymes (3rd edition) 13:1-50(1976).

[2] Fabry S., Lang J., Niermann T., Vingron M., Hensel R. Eur. J. Biochem. 179:405-413(1989).

[3] Zhao G., Pease A.J., Bharani N., Winkler M.E. J. Bacteriol. 177:2804-2812(1995).

271. Granulins signature

Granulins [1] are a family of cysteine-rich peptides of about 6 Kd which may have multiple biological activity. A precursor protein (known as acrogranin) potentially encodes seven different forms of granulin (*grnA* to *grnG*) which are probably released by post-translational proteolytic processing. A schematic representation of the structure of a granulin is shown below: xxxCxxxxxCxxxxxCxxxxxxxxxCxxxxxCxxxxxCxxxxxCxxxxxCx
*****'C': conserved cysteine probably involved in a disulfide bond.'*': position of the pattern. Granulins are evolutionary related to a PMP-D1, a peptide extracted from the pars intercerebralis of migratory locusts [2].

Consensus pattern: C-x-D-x(2)-H-C-C-P-x(4)-C [The four C's are probably involved in disulfide bonds]-

[1] Bhandari V., Palfree R.G., Bateman A. Proc. Natl. Acad. Sci. U.S.A. 89:1715-1719(1992).

[2] Nakakura N., Hietter H., van Dorsselaer A., Luu B. Eur. J. Biochem. 204:147-153(1992).

272. (HCV RdRp) Hepatitis C virus RNA dependent RNA polymerase

The RNA dependent RNA polymerase is also known as non-structural protein NS5B. NS5B is a 65 kDa protein

that resembles other viral RNA polymerases. HCV replication is thought to occur in membrane bound replication complexes. These complexes transcribe the positive strand and the resulting minus strand is used as a template for the synthesis of genomic RNA. There are two viral proteins involved in the reaction, NS3 and NS5B.[1,2]

[1] Lohmann V, Korner F, Herian U, Bartenschlager R; J Virol 1997;71:8416-8428. [2] Behrens SE, Tomei L, De Francesco R; EMBO J 1996;15:12-22. [3] Ishido S, Fujita T, Hotta H; Biochem Biophys Res Commun 1998;244:35-40.

273. (HHH) Helix-hairpin-helix motif.

[1] Doherty AJ, Serpell LC, Ponting CP; Nucleic Acids Res 1996;24:2488-2497.

274. HIT family signature

Recently a family of small proteins of about 12 to 16 Kd has been described[1]. This family currently consists of: - Mammalian protein HINT (also known as Protein kinase C inhibitor 1 or PKCI- 1). HINT was incorrectly thought to be a specific inhibitor of PKC. It has been shown to bind zinc. - Fission yeast diadenosine 5',5'''-P1,P4-tetraphosphate asymmetrical hydrolase (Ap4Aase) (EC 3.6.1.17) [2] (gene aph1), which cleaves A-5'-PPPP- 5'A to yield AMP and ATP. - FHIT, a human protein whose gene is altered in different tumors and which acts [3] as a diadenosine 5',5'''-P1,P3-triphosphate hydrolase (Ap3Aase) (EC 3.6.1.29) cleaving A-5'-PPP-5'A to yield AMP and ADP. - Yeast proteins HNT1 and HNT2. - Maize zinc-binding protein ZBP14. - Escherichia coli hypothetical protein ycfF. - Haemophilus influenzae hypothetical protein HI0961. - Helicobacter pylori hypothetical protein HP0404. - Methanococcus jannaschii hypothetical protein MJ0866. - Mycobacterium leprae hypothetical protein U296A. - Synechocystis strain PCC 6803 hypothetical protein slr1234. - Caenorhabditis elegans hypothetical protein F21C3.3. - A hypothetical 13.2 Kd protein in hisE 3'region in Azospirillum brasilense. - A hypothetical 13.1 Kd protein in p37 5'region in

Mycoplasma hyorhinis. - A hypothetical 12.4 Kd protein in *psbAII* 5'region in *Synechococcus* strain PCC 7942. All these proteins contains a region with three clustered histidines. This region is responsible for the designation of this family: HIT, for 'Histidine Triad' [1]. This region was originally thought to be implied in the binding of a zinc ion but was later identified [4] as part of the alpha-phosphate binding site of a nucleotide-binding domain. As a signature pattern, the region of the histidine triad was selected.

Consensus pattern: [NQA]-x(4)-[GAV]-x-[QF]-x-[LIVM SEQ ID NO:4)]-x-H-[LIVMFYT SEQ ID NO:143)]-H-[LIVMFT SEQ ID NO:282)]-H-[LIVMF SEQ ID NO:2)](2)-[PSGA SEQ ID NO:305)]-

[1] Seraphin B. DNA Seq. 3:177-179(1992).

[2] Huang Y., Garrison P.N., Barnes L.D. Biochem. J. 312:925-932(1995).

[3] Barnes L.D., Garrison P.N., Siprashvili Z., Guranowski A., Robinson A.K., Ingram S.W., Croce C.M., Ohta M., Huebner K. Biochemistry 35:11529-11535(1996).

[4] Brenner C., Garrison P., Gilmour J., Peisach D., Ringe D., Petsko G.A., Lowenstein J.M. Nat. Struct. Biol. 4:231-238(1997).

275. Myc-type, 'helix-loop-helix' dimerization domain signature (HLH)

A number of eukaryotic proteins, which probably are sequence specific DNA-binding proteins that act as transcription factors, share a conserved domain of 40 to 50 amino acid residues. It has been proposed [1] that this domain is formed of two amphipathic helices joined by a variable length linker region that could form a loop. This 'helix-loop-helix' (HLH) domain mediates protein dimerization and has been found in the proteins listed below [2,3,E1,E2]. Most of these proteins have an extra basic region of about 15 amino acid residues that is adjacent to the HLH domain and specifically binds to DNA. They are referred as basic helix-loop-helix proteins (bHLH), and are classified in two groups: class A (ubiquitous) and class B (tissue-specific). Members of the bHLH family bind variations on the core sequence 'CANNTG', also referred to as the E-box motif. The homo- or heterodimerization mediated by the HLH domain is independent of, but necessary for DNA binding, as two basic regions are required for DNA binding activity. The HLH proteins lacking the basic domain (Emc, Id) function as negative regulators since they form

heterodimers, but fail to bind DNA. The hairy-related proteins (hairy, E(spl), deadpan) also repress transcription although they can bind DNA. The proteins of this subfamily act together with co-repressor proteins, like groucho, through their C-terminal motif WRPW. - The myc family of cellular oncogenes [4], which is currently known to contain four members: c-myc [E3], N-myc, L-myc, and B-myc. The myc genes are thought to play a role in cellular differentiation and proliferation. - Proteins involved in myogenesis (the induction of muscle cells). In mammals MyoD1 (Myf-3), myogenin (Myf-4), Myf-5, and Myf-6 (Mrf4 or herculin), in birds CMD1 (QMF-1), in *Xenopus* MyoD and MF25, in *Caenorhabditis elegans* CeMyoD, and in *Drosophila nautilus* (nau). - Vertebrate proteins that bind specific DNA sequences ('E boxes') in various immunoglobulin chains enhancers: E2A or ITF-1 (E12/pan-2 and E47/pan-1), ITF-2 (tcf4), TFE3, and TFEB. - Vertebrate neurogenic differentiation factor 1 that acts as differentiation factor during neurogenesis. - Vertebrate MAX protein, a transcription regulator that forms a sequence-specific DNA-binding protein complex with myc or mad. - Vertebrate Max Interacting Protein 1 (MXI1 protein) which acts as a transcriptional repressor and may antagonize myc transcriptional activity by competing for max. - Proteins of the bHLH/PAS superfamily which are transcriptional activators. In mammals, AH receptor nuclear translocator (ARNT), single-minded homologs (SIM1 and SIM2), hypoxia-inducible factor 1 alpha (HIF1A), AH receptor (AHR), neuronal pas domain proteins (NPAS1 and NPAS2), endothelial pas domain protein 1 (EPAS1), mouse ARNT2, and human BMAL1. In *drosophila*, single-minded (SIM), AH receptor nuclear translocator (ARNT), trachealess protein (TRH), and similar protein (SIMA). - Mammalian transcription factors HES, which repress transcription by acting on two types of DNA sequences, the E box and the N box. - Mammalian MAD protein (max dimerizer) which acts as transcriptional repressor and may antagonize myc transcriptional activity by competing for max. - Mammalian Upstream Stimulatory Factor 1 and 2 (USF1 and USF2), which bind to a symmetrical DNA sequence that is found in a variety of viral and cellular promoters. - Human lyl-1 protein; which is involved, by chromosomal translocation, in T- cell leukemia. - Human transcription factor AP-4. - Mouse helix-loop-helix proteins MATH-1 and MATH-2 which activate E box- dependent transcription in collaboration with E47. - Mammalian stem cell protein (SCL) (also known as tal1), a protein which may play an important role in hemopoietic differentiation. SCL is involved, by chromosomal translocation, in stem-cell leukemia. - Mammalian proteins Id1 to Id4 [5]. Id (inhibitor of DNA binding) proteins lack a basic DNA-binding domain but are able to form heterodimers with other HLH proteins,

thereby inhibiting binding to DNA. - *Drosophila* extra-macrochaetae (emc) protein, which participates in sensory organ patterning by antagonizing the neurogenic activity of the achaete-scute complex. Emc is the homolog of mammalian Id proteins. - Human Sterol Regulatory Element Binding Protein 1 (SREBP-1), a transcriptional activator that binds to the sterol regulatory element 1 (SRE-1) found in the flanking region of the LDLR gene and in other genes. - *Drosophila* achaete-scute (AS-C) complex proteins T3 (l'sc), T4 (scute), T5 (achaete) and T8 (asense). The AS-C proteins are involved in the determination of the neuronal precursors in the peripheral nervous system and the central nervous system. - Mammalian homologs of achaete-scute proteins, the MASH-1 and MASH-2 proteins. - *Drosophila* atonal protein (ato) which is involved in neurogenesis. - *Drosophila* daughterless (da) protein, which is essential for neurogenesis and sex-determination. - *Drosophila* deadpan (dnp), a hairy-like protein involved in the functional differentiation of neurons. - *Drosophila* delilah (dei) protein, which plays an important role in the differentiation of epidermal cells into muscle. - *Drosophila* hairy (h) protein, a transcriptional repressor which regulates the embryonic segmentation and adult bristle patterning. - *Drosophila* enhancer of split proteins E(spl), that are hairy-like proteins active during neurogenesis. also act as transcriptional repressors. - *Drosophila* twist (twi) protein, which is involved in the establishment of germ layers in embryos. - Maize anthocyanin regulatory proteins R-S and LC. - Yeast centromere-binding protein 1 (CPF1 or CBF1). This protein is involved in chromosomal segregation. It binds to a highly conserved DNA sequence, found in centromeres and in several promoters. - Yeast INO2 and INO4 proteins. - Yeast phosphate system positive regulatory protein PHO4 which interacts with the upstream activating sequence of several acid phosphatase genes. - Yeast serine-rich protein TYE7 that is required for ty-mediated ADH2 expression. - *Neurospora crassa* nuc-1, a protein that activates the transcription of structural genes for phosphorus acquisition. - Fission yeast protein esc1 which is involved in the sexual differentiation process. The schematic representation of the helix-loop-helix domain is shown here: xxxxxxxxxxxxxxxxxxxxxxxxxxxx-----xxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Amphipathic helix 1 Loop Amphipathic helix 2. The signature pattern developed to detect this domain spans completely the second amphipathic helix.

Consensus pattern: [DENSTAP SEQ ID NO:306)]-[KTR)]-[LIVMAGSNT SEQ ID NO:307)]-{FYWCPHKR SEQ ID NO:308)}-[LIVMT SEQ ID NO:1)]-[LIVM SEQ ID NO:4)]- x(2)-[STAV SEQ ID NO:105)]-[LIVMSTACKR SEQ ID NO:309)]-x-[VMFYH

SEQ ID NO:310)]-[LIVMTA SEQ ID NO:311)]-{P}-{P}-[LIVMRKHQ SEQ ID NO:312)].-

[1] Murre C., McCaw P.S., Baltimore D. Cell 56:777-783(1989).

5 [2] Garrel J., Campuzano S. BioEssays 13:493-498(1991).

[3] Kato G.J., Dang C.V. FASEB J. 6:3065-3072(1992).

[4] Krause M., Fire A., Harrison S.W., Priess J., Weintraub H. Cell 63:907-919(1990).

[5] Riechmann V., van Cruechten I., Sablitzky F. Nucleic Acids Res. 22:749-755(1994).

10

276. HMG14 and HMG17 signature

High mobility group (HMG) proteins are a family of relatively low molecular weight non-histone components in chromatin. HMG14 and HMG17 [1], two related proteins of about 100 amino acid residues, bind to the inner side of the nucleosomal DNA thus altering the interaction between the DNA and the histone octamer. These two proteins may be involved in the process which maintains transcribable genes in a unique chromatin conformation. The trout nonhistone chromosomal protein H6 (histone T) also belongs to this family. As a signature pattern a conserved stretch of 10 residues located in the N-terminal section of HMG14 and HMG17 was selected.

20

Consensus pattern: R-R-S-A-R-L-S-A-[RK]-P-

[1] Bustin M., Reeves R. Prog. Nucleic Acid Res. Mol. Biol. 54:35-100(1996).

25

277. Hydroxymethylglutaryl-coenzyme A lyase active site (HMGL1)

3-hydroxy-3-methylglutaryl-coenzyme A lyase (HMG-CoA lyase or HL) (EC 4.1.3.4)catalyzes the transformation of HMG-CoA into acetyl-CoA and acetoacetate. In vertebrates it is a mitochondrial enzyme which is involved in ketogenesis and in leucine catabolism [1]. In some bacteria, such as *Pseudomonas mevalonii*, it is involved in mevalonate catabolism (gene *mvaB*). A cysteine has been shown[2], in *mvaB*, to be required for the activity of the enzyme. The region around this residue is perfectly conserved and is used as a signature pattern.

30

Consensus pattern: S-V-A-G-L-G-G-C-P-Y [C is the active site residue]-

[1] Mitchell G.A., Robert M.-F., Hruz P.W., Wang S., Fontaine G., Behnke C.E., Mende-
Mueller L.M., Schappert K., Lee C., Gibson K.M., Miziorko H.M. J. Biol. Chem. 268:4376-
4381(1993).

[2] Hruz P.W., Narasimhan C., Miziorko H.M. Biochemistry 31:6842-6847(1992).

Alpha-isopropylmalate and homocitrate synthases signatures (HMGL2)

The following enzymes have been shown [1] to be functionally as well as evolutionary
related: - Alpha-isopropylmalate synthase (EC 4.1.3.12) which catalyzes the first step in the
biosynthesis of leucine, the condensation of acetyl-CoA and alpha- ketoisovalerate to form 2-
isopropylmalate synthase. - Homocitrate synthase (EC 4.1.3.21) (gene nifV) which is
involved in the biosynthesis of the iron-molybdenum cofactor of nitrogenase and catalyzes
the condensation of acetyl-CoA and alpha-ketoglutarate into homocitrate. - Soybean late
nodulin 56. - Methanococcus jannaschii hypothetical proteins MJ0503, MJ1195 and MJ1392.
Two conserved regions were selected as signature patterns for these enzymes. The first region
is located in the N-terminal section while the second region is located in the central section
and contains two conserved histidine residues which could be implicated in the catalytic
mechanism.

Consensus pattern: L-R-[DE]-G-x-Q-x(10)-K-

Consensus pattern: [LIVMFW SEQ ID NO:13])-x(2)-H-x-H-[DN]-D-x-G-x-[GAS]-x-
[GASLI SEQ ID NO:313)]-

[1] Wang S.-Z., Dean D.R., Chen J.-S., Johnson J.L. J. Bacteriol. 173:3041-3046(1991).

278. (HMG C0A synt) Hydroxymethylglutaryl-coenzyme A synthase active site

Hydroxymethylglutaryl-coenzyme A synthase (EC 4.1.3.5) (HMG-CoA synthase) catalyzes
the condensation of acetyl-CoA with acetoacetyl-CoA to produce HMG- CoA and CoA [1].In
vertebrates there are two isozymes located in different subcellular compartments: a cytosolic
form which is the starting point of the mevalonate pathway which leads to cholesterol and

other sterolic and isoprenoid compounds and a mitochondrial form responsible for ketone body biosynthesis. HMG-CoA is also found in other eukaryotes such as insect, plants and fungi. A cysteine is known to act as the catalytic nucleophile in the first step of the reaction, the acetylation of the enzyme by acetyl-CoA. The conserved region was used around this active site residue as a signature pattern.

Consensus pattern: N-x-[DN]-[IV]-E-G-[IV]-D-x(2)-N-A-C-[FY]-x-G [C is the active site residue]-

[1] Rokosz L.L., Boulton D.A., Butkiewicz E.A., Sanyal G., Cueto M.A., Lachance P.A., Hermes J.D. Arch. Biochem. Biophys. 312:1-13(1994).

279. HMG (high mobility group) box

280. HSF-type DNA-binding domain signature

Heat shock factor (HSF) is a DNA-binding protein that specifically binds heat shock promoter elements (HSE). HSE is a palindromic element rich with repetitive purine and pyrimidine motifs: 5'-nGAAnnTTCnnGAAnnTTCn-3'. HSF is expressed at normal temperatures but is activated by heat shock or chemical stressors [1,2]. The sequences of HSF from various species show extensive similarity in a region of about 90 amino acids, which has been shown [3] to bind DNA. Some other proteins also contain a HSF domain, these are:

- Yeast SFL1, a protein involved in cell surface assembly and regulation of the gene related to flocculation (asexual cell aggregation) [4].
- Yeast transcription factor SKN7 (or BRY1 or POS9), which binds to the promoter elements SCB and MCB essential for the control of G1 cyclins expression [5].
- Yeast MGA1.
- Yeast hypothetical protein YJR147w.

A pattern from the most conserved part of the HSF DNA-binding domain was derived, its central region.

Consensus pattern: L-x(3)-[FY]-K-H-x-N-x-[STAN SEQ ID NO:250)]-S-F-[LIVM SEQ ID NO:4)]-R-Q-L-[NH]-x-Y-x- [FYW]-[RKH]-K-[LIVM SEQ ID NO:4)]-

[1] Sorger P.K. Cell 65:363-366(1991).

[2] Mager W.H., Moradas Ferreira P. Biochem. J. 290:1-13(1993).

[3] Vuister G.W., Kim S.-J., Orosz A., Marquardt J., Wu C., Bax A. Nat. Struct. Biol. 1:605-613(1994).

[4] Fujita A., Kikuchi Y., Kuhara S., Misumi Y., Matsumoto S., Kobayashi H. Gene 85:321-328(1989).

[5] Morgan B.A., Bouquin N., Merrill G.F., Johnston L.H. EMBO J. 14:5679-5689(1995).

281. Heat shock hsp20 proteins family profile

Prokaryotic and eukaryotic organisms respond to heat shock or other environmental stress by inducing the synthesis of proteins collectively known as heat-shock proteins (hsp) [1].

Amongst them is a family of proteins with an average molecular weight of 20 Kd, known as the hsp20 proteins [2 to 5]. These seem to act as chaperones that can protect other proteins against heat-induced denaturation and aggregation. Hsp20 proteins seem to form large

heterooligomeric aggregates; their family is currently composed of the following members: - Vertebrate heat shock protein hsp27 (hsp25), induced by a variety of environmental stresses.

- *Drosophila* heat shock proteins hsp22, hsp23, hsp26, hsp27, hsp67BA and BC. -

Caenorhabditis elegans hsp16 multigene family. - Fungal HSP26 (budding yeast) and hsp30

(*Neurospora crassa* and *Aspergillus Nidulans*). - Plant small hsp's. Plants have four classes of

hsp20: classes I and II which are cytoplasmic, class III which is chloroplastic and class IV which is found in the endomembrane. - Alpha-crystallin A and B chains. Alpha-crystallin is an abundant constituent of the eye lens of most vertebrate species. Its main function appears to be to maintain the correct refractive index of the lens. It is also found in other tissues where it seems to act as a chaperone [6]. - *Schistosoma mansoni* major egg antigen p40.

Structurally, p40 is built of two tandem hsp20 domains. - A variety of prokaryotic proteins: ibpA and ibpB from *Escherichia coli*, hsp18 from *Clostridium acetobutylicum*, spore protein SP21 (hspA) from *Stigmatella aurantiaca*, *Mycobacterium leprae* 18 Kd antigen and *Mycobacterium tuberculosis* 14 Kd antigen. - *Methanococcus jannaschii* hypothetical protein MJ0285. Structurally, this family is characterized by the presence of a conserved C-terminal domain of about 100 residues. The profile developed to detect members of the hsp20 family is based on an alignment of this domain.

-Sequences known to belong to this class detected by the profile: ALL.

[1] Lindquist S., Craig E.A. *Annu. Rev. Genet.* 22:631-677(1988).[2] de Jong W.W., Leunissen J.A.M., Voorter C.E.M. *Mol. Biol. Evol.* 10:103-126(1993).[3] Caspers G.J., Leunissen J.A.M., de Jong W.W. *J. Mol. Evol.* 40:238-248(1995).[4] Jaenicke R., Creighton T.E. *Curr. Biol.* 3:234-235(1993).[5] Jakob U., Buchner J. *Trends Biochem. Sci.* 19:205-211(1994).[6] Groenen P.J.T.A., Merck K.B., de Jong W.W., Bloemendal H. *Eur. J. Biochem.* 225:1-9(1994).

282. Heat shock hsp70 proteins family signatures

Prokaryotic and eukaryotic organisms respond to heat shock or other environmental stress by the induction of the synthesis of proteins collectively known as heat-shock proteins (hsp) [1]. Amongst them is a family of proteins with an average molecular weight of 70 Kd, known as the hsp70proteins [2,3,4]. In most species, there are many proteins that belong to the hsp70 family. Some of them are expressed under unstressed conditions. Hsp70proteins can be found in different cellular compartments (nuclear, cytosolic, mitochondrial, endoplasmic reticulum, etc.). Some of the hsp70 family proteins are listed below: - In *Escherichia coli* and other bacteria, the main hsp70 protein is known as the dnaK protein. A second protein, hscA, has been recently discovered. dnaK is also found in the chloroplast genome of red algae. - In yeast, at least ten hsp70 proteins are known to exist: SSA1 to SSA4, SSB1, SSB2, SSC1, SSD1 (KAR2), SSE1 (MSI3) and SSE2. - In *Drosophila*, there are at least eight different hsp70 proteins: HSP70, HSP68, and HSC-1 to HSC-6. - In mammals, there are at least eight different proteins: HSPA1 to HSPA6, HSC70, and GRP78 (also known as the immunoglobulin heavy chain binding protein (BiP)). - In the sugar beet yellow virus (SBYV), a hsp70 homolog has been shown [5] to exist. - In archaeobacteria, hsp70 proteins are also present [6]. All proteins belonging to the hsp70 family bind ATP. A variety of functions has been postulated for hsp70 proteins. It now appears [7] that some hsp70proteins play an important role in the transport of proteins across membranes. They also seem to be involved in protein folding and in the assembly/disassembly of protein complexes [8]. Three signature patterns for the hsp70 family of proteins were derived; the first centered on a conserved pentapeptide found in the N-terminal section of these proteins; the two others on conserved regions located in the central part of the sequence.

Consensus pattern: [IV]-D-L-G-T-[ST]-x-[SC] -

Consensus pattern: [LIVMF SEQ ID NO:2)]-[LIVMFY SEQ ID NO:18)]-[DN]-[LIVMFS
SEQ ID NO:132)]-G-[GSH]-[GS]-[AST]-x(3)- [ST]-[LIVM SEQ ID NO:4)]-[LIVMFC SEQ
ID NO:90)]-

Consensus pattern: [LIVMY SEQ ID NO:141)]-x-[LIVMF SEQ ID NO:2)]-x-G-G-x-[ST]-x-
5 [LIVM SEQ ID NO:4)]-P-x-[LIVM SEQ ID NO:4)]-x- [DEQKRSTA SEQ ID NO:314)]-

[1] Lindquist S., Craig E.A. Annu. Rev. Genet. 22:631-677(1988).

[2] Pelham H.R.B. Cell 46:959-961(1986).

[3] Pelham H.R.B. Nature 332:776-77(1988).[4] Craig E.A. BioEssays 11:48-52(1989).

10 [5] Agranovsky A.A., Boyko V.P., Karasev A.V., Koonin E.V., Dolja V.V. J. Mol. Biol.
217:603-610(1991).

[6] Gupta R.S., Singh B. J. Bacteriol. 174:4594-4605(1992).

[7] Deshaies R.J., Koch B.D., Schekmam R. Trends Biochem. Sci. 13:384-388(1988).

[8] Craig E.A., Gross C.A. Trends Biochem. Sci. 16:135-140(1991).

15

283. Heat shock hsp90 proteins family signature

Prokaryotic and eukaryotic organisms respond to heat shock or other environmental stress by
the induction of the synthesis of proteins collectively known as heat-shock proteins (hsp) [1].

20 Amongst them is a family of proteins, with an average molecular weight of 90 Kd, known as
the hsp90proteins. Proteins known to belong to this family are: - Escherichia coli and other
bacteria heat shock protein c62.5 (gene htpG). - Vertebrate hsp 90-alpha (hsp 86) and hsp 90-
beta (hsp 84). - Drosophila hsp 82 (hsp 83). - Trypanosoma cruzi hsp 85. - Plants Hsp82 or
Hsp83. - Yeast and other fungi HSC82, and HSP82. - The endoplasmic reticulum protein
25 'endoplasmin' (also known as Erp99 in mouse, GRP94 in hamster, and hsp 108 in
chicken). The exact function of hsp90 proteins is not yet known. In higher eukaryotes, hsp90
has been found associated with steroid hormone receptors, with tyrosine kinase oncogene
products of several retroviruses, with eIF2alpha kinase, and with actin and tubulin. Hsp90 are
probable chaperonins that possess ATPase activity [2,3]. As a signature pattern for the hsp90
30 family of proteins, a highly conserved region found in the N-terminal part of these proteins
was selected.

Consensus pattern: Y-x-[NQH]-K-[DE]-[IVA]-F-L-R-[ED] -

- [1] Lindquist S., Craig E.A. Annu. Rev. Genet. 22:631-677(1988).
[2] Nadeau K., Das A., Walsh C.T. J. Biol. Chem. 268:1479-1487(1993).
[3] Jakob U., Buchner J. Trends Biochem. Sci. 19:205-211(1994).

5

284. Helix-turn-helix (HTH3)

This large family of DNA binding helix-turn helix proteins includes Cro
Swiss:P03036 and CI Swiss:P03034.

10

285. Heme oxygenase signature

Heme oxygenase (EC 1.14.99.3) (HO) [1] is the microsomal enzyme that, in animals, carries out the oxidation of heme, it cleaves the heme ring at the alpha methene bridge to form biliverdin and carbon monoxide. Biliverdin is subsequently converted to bilirubin by biliverdin reductase. In mammals there are three isozymes of heme oxygenase: HO-1 to HO-3. The first two isozymes differ in their tissue expression and their inducibility: HO-1 is highly inducible by its substrate heme and by various non-heme substances, while HO-2 is non-inducible. It has been suggested [2] that HO-2 could be implicated in the production of carbon monoxide in the brain where it is said to act as a neurotransmitter. In the genome of the chloroplast of red algae as well as in cyanobacteria, there is a heme oxygenase (gene pbsA) that is the key enzyme in the synthesis of the chromophoric part of the photosynthetic antennae [3]. An heme oxygenase is also present in the bacteria *Corynebacterium diphtheriae* (gene hmuO), where it is involved in the acquisition of iron from the host heme [4]. There is, in the central section of these enzymes, a well conserved region centered on a histidine residue which is proposed to play a key role in binding the substrate heme at the active center of the enzyme. This region was used as a signature pattern.

15

20

25

Consensus pattern: L-[IV]-A-H-[STACH SEQ ID NO:315)]-Y-[STV]-[RT]-Y-[LIVM SEQ ID NO:4)]-G [H binds the heme] -

30

- [1] Maines M.D. FASEB J. 2:2557-2568(1988).
[2] Barinaga M. Science 259:309-309(1993).

[3] Richaud C., Zabulon G. Proc. Natl. Acad. Sci. U.S.A. 94:11736-11741(1997).

[4] Schmitt M.P. J. Bacteriol. 179:838-845(1997).

5 286. Hepatitis core antigen.

The core antigen of hepatitis viruses possesses a carboxyl terminus rich in arginine. On this basis it was predicted that the core antigen would bind DNA [1]. There is some
10 experimental evidence to support this [2].

[1] Pasek M, Goto T, Gilbert W, Zink B, Schaller H, Mckay P, Leadbetter G, Murray K; Nature 1979;282:575-579. [2] Gallina A, Bonelli F, Zentilin L, Rindi G, Muttini M,
15 Milanesi G; J Virol 1989;63:4645-4652.

287. Histidine biosynthesis protein

Proteins involved in steps 4 and 6 of the histidine biosynthesis pathway are contained
20 in this family. Histidine is formed by several complex and distinct biochemical reactions catalysed by eight enzymes. The enzymes in this Pfam entry are called His6 and His7 in eukaryotes and HisA and HisF in prokaryotes.

[1] Fani R, Tamburini E, Mori E, Lazcano A, Lio P, Barberio C, Casalone E, Cavalieri D, Perito B, Polsinelli M, Gene 1997;197:9-17. [2] Fani R, Lio P, Chiarelli I,
25 Bazzicalupo M, J Mol Evol 1994;38:489-495.

288. Histone deacetylase family

Histones can be reversibly acetylated on several lysine residues. Regulation of
30 transcription is caused in part by this mechanism. Histone deacetylases catalyse the removal of the acetyl group. Histone deacetylases are related to other proteins [1].

Leipe DD, Landsman D, Nucleic Acids Res 1997;25:3693-3697.

289. Histidinol dehydrogenase signature

Histidinol dehydrogenase (EC 1.1.1.23) (HDH) catalyzes the terminal step in the biosynthesis of histidine in bacteria, fungi, and plants, the four-electron oxidation of L-histidinol to histidine. In bacteria HDH is a single chain polypeptide; in fungi it is the C-terminal domain of a multifunctional enzyme which catalyzes three different steps of histidine biosynthesis; and in plants it is expressed as nuclear encoded protein precursor which is exported to the chloroplast [1]. As a signature pattern a highly conserved region located in the central part of HDH was selected. This region does not correspond to the part of the enzyme that, in most, but not all HDH sequences contains a cysteine residue which, in *Salmonella typhimurium*, has been said [2] to be important for the catalytic activity of the enzyme.

Consensus pattern: I-D-x(2)-A-G-P-[ST]-E-[LIVS SEQ ID NO:316)]-[LIVMA SEQ ID NO:30)](3)-[AC]-x(3)-A-x(4)-[LIVM SEQ ID NO:4)]-[AV]-[SACL SEQ ID NO:317)]-[DE]-[LIVMFC SEQ ID NO:90)]-[LIVM SEQ ID NO:4)]-[SA]-x(2)-E-H-

[1] Nagai A., Ward E., Beck J., Tada S., Chang J.-Y., Scheidegger A., Ryals J. Proc. Natl. Acad. Sci. U.S.A. 88:4133-4137(1991).

[2] Grubmeyer C.T., Gray W.R. Biochemistry 25:4778-4784(1986).

290. Homoserine dehydrogenase signature

Homoserine dehydrogenase (EC 1.1.1.3) (HDh) [1,2] catalyzes NAD-dependent reduction of aspartate beta-semialdehyde into homoserine. This reaction is the third step in a pathway leading from aspartate to homoserine. The latter participates in the biosynthesis of threonine and then isoleucine as well as in that of methionine. HDh is found either as a single chain protein as in some bacteria and yeast, or as a bifunctional enzyme consisting of an N-terminal aspartokinase domain and a C-terminal HDh domain as in bacteria such as *Escherichia coli* and in plants. As a signature pattern, the best conserved region of Hdh has been selected. This is a segment of 23 to 24 residues located in the central section and that contains two conserved aspartate residues.

Consensus pattern: A-x(3)-G-[LIVMFY SEQ ID NO:18)]-[STAG SEQ ID NO:20)]-x(2,3)-[DNS]-P-x(2)-D-[LIVM SEQ ID NO:4)]-x-G- x-D-x(3)-K-

[1] Thomas D., Barbey R., Surdin-Kerjan Y. FEBS Lett. 323:289-293(1993).

5 [2] Cami B., Clepet C., Patte J.-C. Biochimie 75:487-495(1993).

291. haloacid dehalogenase-like hydrolase

This family is structurally different from the alpha/ beta hydrolase family (abhydrolase). This family includes L-2-haloacid dehalogenase, epoxide hydrolases and phosphatases. The structure of the family consists of two domains. One is an inserted four helix bundle, which is the least well conserved region of the alignment, between residues 16 and 96 of [Swiss:P24069](#). The rest of the fold is composed of the core alpha/beta domain.

[1] Hisano T, Hata Y, Fujii T, Liu JQ, Kurihara T, Esaki N, Soda K, J Biol Chem 1996; 271:20322-20330.

292. DEAD and DEAH box families ATP-dependent helicases signatures (helicase_C)

A number of eukaryotic and prokaryotic proteins have been characterized [1,2,3] on the basis of their structural similarity. They all seem to be involved in ATP-dependent, nucleic-acid unwinding. Proteins currently known to belong to this family are: - Initiation factor eIF-4A. Found in eukaryotes, this protein is a subunit of a high molecular weight complex involved in 5'cap recognition and the binding of mRNA to ribosomes. It is an ATP-dependent RNA-helicase. - PRP5 and PRP28. These yeast proteins are involved in various ATP-requiring steps of the pre-mRNA splicing process. - P110, a mouse protein expressed specifically during spermatogenesis. - An3, a Xenopus putative RNA helicase, closely related to P110. - SPP81/DED1 and DBP1, two yeast proteins probably involved in pre-mRNA splicing and related to P110. - Caenorhabditis elegans helicase glh-1. - MSS116, a yeast protein required for mitochondrial splicing. - SPB4, a yeast protein involved in the maturation of 25S ribosomal RNA. - p68, a human nuclear antigen. p68 has ATPase and DNA-helicase activities in vitro. It is involved in cell growth and division. - Rm62 (p62), a Drosophila putative RNA helicase related to p68. - DBP2, a yeast protein related to p68. - DHH1, a yeast protein. - DRS1, a yeast protein involved in ribosome assembly. - MAK5, a yeast protein

involved in maintenance of dsRNA killer plasmid. - ROK1, a yeast protein. - ste13, a fission yeast protein. - Vasa, a Drosophila protein important for oocyte formation and specification of embryonic posterior structures. - Me31B, a Drosophila maternally expressed protein of unknown function. - dbpA, an Escherichia coli putative RNA helicase. - deaD, an Escherichia coli putative RNA helicase which can suppress a mutation in the rpsB gene for ribosomal protein S2. - rhlB, an Escherichia coli putative RNA helicase. - rhlE, an Escherichia coli putative RNA helicase. - srmB, an Escherichia coli protein that shows RNA-dependent ATPase activity. It probably interacts with 23S ribosomal RNA. - Caenorhabditis elegans hypothetical proteins T26G10.1, ZK512.2 and ZK686.2. - Yeast hypothetical protein YHR065c. - Yeast hypothetical protein YHR169w. - Fission yeast hypothetical protein SpAC31A2.07c. - Bacillus subtilis hypothetical protein yxiN. All these proteins share a number of conserved sequence motifs. Some of them are specific to this family while others are shared by other ATP-binding proteins or by proteins belonging to the helicases 'superfamily' [4,E1]. One of these motifs, called the 'D-E-A-D-box', represents a special version of the B motif of ATP-binding proteins. Some other proteins belong to a subfamily which have His instead of the second Asp and are thus said to be 'D-E-A-H-box' proteins [3,5,6,E1]. Proteins currently known to belong to this subfamily are: - PRP2, PRP16, PRP22 and PRP43. These yeast proteins are all involved in various ATP-requiring steps of the pre-mRNA splicing process. - Fission yeast prh1, which may be involved in pre-mRNA splicing. - Male-less (mle), a Drosophila protein required in males, for dosage compensation of X chromosome linked genes. - RAD3 from yeast. RAD3 is a DNA helicase involved in excision repair of DNA damaged by UV light, bulky adducts or cross-linking agents. Fission yeast rad15 (rhp3) and mammalian DNA excision repair protein XPD (ERCC-2) are the homologs of RAD3. - Yeast CHL1 (or CTF1), which is important for chromosome transmission and normal cell cycle progression in G(2)/M. - Yeast TPS1. - Yeast hypothetical protein YKL078w. - Caenorhabditis elegans hypothetical proteins C06E1.10 and K03H1.2. - Poxviruses' early transcription factor 70 Kd subunit which acts with RNA polymerase to initiate transcription from early gene promoters. - I8, a putative vaccinia virus helicase. - hrpA, an Escherichia coli putative RNA helicase. Signature patterns were developed for both subfamilies.

Consensus pattern: [LIVMF SEQ ID NO:2]](2)-D-E-A-D-[RKEN SEQ ID NO:196]]-x-[LIVMFYGSTN SEQ ID NO:197]]-

Consensus pattern: [GSAH SEQ ID NO:198])-x-[LIVMF SEQ ID NO:2)](3)-D-E-[ALIV SEQ ID NO:199)]-H-[NECR SEQ ID NO:200)] -

Note: proteins belonging to this family also contain a copy of the ATP/GTP- binding motif 'A' (P-loop) (see the relevant entry <[PDOC00017](#)

5

[1] Schmid S.R., Linder P. Mol. Microbiol. 6:283-292(1992).

[2] Linder P., Lasko P., Ashburner M., Leroy P., Nielsen P.J., Nishi K., Schnier J., Slonimski P.P. Nature 337:121-122(1989).

[3] Wassarman D.A., Steitz J.A. Nature 349:463-464(1991).

10 [4] Hodgman T.C. Nature 333:22-23(1988) and Nature 333:578-578(1988) (Errata).

[5] Harosh I., Deschavanne P. Nucleic Acids Res. 19:6331-6331(1991).

[6] Koonin E.V., Senkevich T.G. J. Gen. Virol. 73:989-993(1992).

15 293. Heme-binding domain in cytochrome b5 and oxidoreductases (heme_1)

Cytochrome b5 is a membrane-bound hemo protein which acts as an electron carrier for several membrane-bound oxygenases [1]. There are two homologous forms of b5, one found in microsomes and one found in the outer membrane of mitochondria. Two conserved histidine residues serve as axial ligands for the heme group. The structure of a number of oxidoreductases consists of the juxtaposition of a heme-binding domain homologous to that of b5 and either a flavodehydrogenase or a molybdopterin domain. These enzymes are:

- Lactate dehydrogenase (EC [1.1.2.3](#)) [2], an enzyme that consists of a flavodehydrogenase domain and a heme-binding domain called cytochrome b2.
- Nitrate reductase (EC [1.6.6.1](#)), a key enzyme involved in the first step of nitrate assimilation in plants, fungi and bacteria [3,4]. Consists of a molybdopterin domain (see <[PDOC00484](#)>), a heme-binding domain called cytochrome b557, as well as a cytochrome reductase domain.
- Sulfite oxidase (EC [1.8.3.1](#)) [5], which catalyzes the terminal reaction in the oxidative degradation of sulfur-containing amino acids. Also consists of a molybdopterin domain and a heme-binding domain.

30

This family of proteins also includes:

- TU-36B, a Drosophila muscle protein of unknown function [6].
- Fission yeast hypothetical protein SpAC1F12.10c.

299

- Yeast hypothetical protein YMR073c.
- Yeast hypothetical protein YMR272c.

A segment was used which includes the first of the two histidine heme ligands, as a signature pattern for the heme-binding domain of cytochrome b5 family.

5

Consensus pattern: [FY]-[LIVMK SEQ ID NO:281)]-x(2)-H-P-[GA]-G [H is a heme axial ligand]-

[1] Ozols J. Biochim. Biophys. Acta 997:121-130(1989).

10 [2] Guiard B. EMBO J. 4:3265-3272(1985).

[3] Calza R., Huttner E., Vincentz M., Rouze P., Galangau F., Vaucheret H., Cherel I., Meyer C., Kronenberger J., Caboche M. Mol. Gen. Genet. 209:552-562(1987).

[4] Crawford N.M., Smith M., Bellissimo D., Davis R.W. Proc. Natl. Acad. Sci. U.S.A. 85:5006-5010(1988).

15 [5] Guiard B., Lederer F. Eur. J. Biochem. 100:441-453(1979).

[6] Levin R.J., Boychuk P.L., Croniger C.M., Kazzaz J.A., Rozek C.E. Nucleic Acids Res. 17:6349-6367(1989).

20 294. Hexapeptide-repeat containing-transferases signature

On the basis of sequence similarity, a number of transferases have been proposed [1,2,3,4] to belong to a single family. These proteins are: - Serine acetyltransferase (EC 2.3.1.30) (SAT) (gene *cysE*), an enzyme involved in cysteine biosynthesis. - *Azotobacter chroococcum* nitrogen fixation protein *nifP*. *NifP* is most probably a SAT involved in the optimization of
 25 nitrogenase activity. - *Escherichia coli* thiogalactoside acetyltransferase (EC 2.3.1.18) (gene *lacA*), an enzyme involved in the biosynthesis of lactose. - UDP-N-acetylglucosamine acyltransferase (EC 2.3.1.129) (gene *lpxA*), an enzyme involved in the biosynthesis of lipid A, a phosphorylated glycolipid that anchors the lipopolysaccharide to the outer membrane of the cell. - UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.-) (gene
 30 *lpxD* or *firA*), which is also involved in the biosynthesis of lipid A. - Chloramphenicol acetyltransferase (CAT) (EC 2.3.1.28) from *Agrobacterium tumefaciens*, *Bacillus sphaericus*, *Escherichia coli* plasmid IncFII NR79, *Pseudomonas aeruginosa*, *Staphylococcus aureus* plasmid pIP630. These CAT are not evolutionary related to the main family of CAT (see

<PDOC00093>). - Rhizobium nodulation protein nodL. NodL is an acetyltransferase involved in the O-acetylation of Nod factors. - Bacterial maltose O-acetyltransferase (EC 2.3.1.79). - Bacterial tetrahydrodipicolinate N-succinyltransferase (EC 2.3.1.117) (gene dapD) which catalyzes the fourth step in the biosynthesis of diaminopimelate and lysine from aspartate semialdehyde. - Bacterial N-acetylglucosamine-1-phosphate uridyltransferase (EC 2.7.7.23) (gene glmU or gcaD or tms), an enzyme involved in peptidoglycan and lipopolysaccharide biosynthesis. - Staphylococcus aureus protein capG which is involved in biosynthesis of type 1 capsular polysaccharide. - Yeast hypothetical protein YJL218w, which is highly similar to Escherichia coli lacA. - Fission yeast hypothetical protein SpAC18B11.09c. - Methanococcus jannaschii hypothetical protein MJ1064. These proteins have been shown [3,4] to contain a repeat structure composed of tandem repeats of a [LIV]-G-x(4) hexapeptide which, in the tertiary structure of lpxA [5], has been shown to form a left-handed parallel beta helix. Our signature pattern is based on a fourfold repeat of this hexapeptide.

Consensus pattern: [LIV]-[GAED SEQ ID NO:318)]-x(2)-[STAV SEQ ID NO:105)]-x-[LIV]-x(3)-[LIVAC SEQ ID NO:319)]-x-[LIV]- [GAED SEQ ID NO:318)]-x(2)-[STAVR SEQ ID NO:320)]-x-[LIV]-[GAED SEQ ID NO:318)]-x(2)-[STAV SEQ ID NO:105)]-x-[LIV]- x(3)-[LIV]-

[1] Downie J.A. Mol. Microbiol. 3:1649-1651(1989).

[2] Parent R., Roy P.H. J. Bacteriol. 174:2891-2897(1992).

[3] Vaara M. FEMS Microbiol. Lett. 97:249-254(1992).

[4] Vuorio R., Haerkonen T., Tolvanen M., Vaara M. FEBS Lett. 337:289-292(1994).

[5] Raetz C.R.H., Roderick S.L. Science 270:997-1000(1995).

295. Hexokinases signature. Hexokinase (EC 2.7.1.1) [1,2] is an important glycolytic enzyme that catalyzes the phosphorylation of keto- and aldohexoses (e.g. glucose, mannose and fructose) using MgATP as the phosphoryl donor. In vertebrates there are four major isoenzymes, commonly referred as types I,II, III and IV. Type IV hexokinase, which is often incorrectly designated glucokinase [3], is only expressed in liver and pancreatic beta-cells and plays an important role in modulating insulin secretion; it is a protein of a molecular

mass of about 50 Kd. Hexokinases of types I to III, which have low Km values for glucose, have a molecular mass of about 100 Kd. Structurally they consist of a very small N-terminal hydrophobic membrane-binding domain followed by two highly similar domains of 450 residues. The first domain has lost its catalytic activity and has evolved into a regulatory domain. In yeast there are three different isozymes: hexokinase PI (gene HXK1), PII (gene HXKB), and glucokinase (gene GLK1). All three proteins have a molecular mass of about 50 Kd. All these enzymes contain one (or two in the case of types I to III isozymes) strongly conserved region which has been shown [4] to be involved in substrate binding. A pattern from that region has been derived

Consensus pattern: [LIVM SEQ ID NO:4)]-G-F-[TN]-F-S-[FY]-P-x(5)-[LIVM SEQ ID NO:4)]-[DNST SEQ ID NO:265)]-x(3)-[LIVM SEQ ID NO:4)]- x(2)-W-T-K-x-[LF]-

[1] Middleton R.J. Biochem. Soc. Trans. 18:180-183(1990).[2] Griffin L.D., Gelb B.D., Wheeler D.A., Davison D., Adams V., McCabe E.R. Genomics 11:1014-1024(1991).[3] Cornish-Bowden A., Luz Cardenas M. Trends Biochem. Sci. 16:281-282(1991).[4] Schirch D.M., Wilson J.E. Arch. Biochem. Biophys. 254:385-396(1987).

296. Histone H2A signature (his1)

Histone H2A is one of the four histones, along with H2B, H3 and H4, which forms the eukaryotic nucleosome core. Using alignments of histone H2A sequences [1,2,E1] as a signature pattern, a conserved region in the N-terminal part of H2A. This region is conserved both in classical S-phase regulated H2A's and in variant histone H2A's which are synthesized throughout the cell cycle.

Consensus pattern: [AC]-G-L-x-F-P-V-

[1] Wells D.E., Brown D. Nucleic Acids Res. 19:2173-2188(1991).

[2] Thatcher T.H., Gorovsky M.A. Nucleic Acids Res. 22:174-179(1994).

Histone H4 signature (his2)

Histone H4 is one of the four histones, along with H2A, H2B and H3, which forms the eukaryotic nucleosome core. Along with H3, it plays a central role in nucleosome formation. The sequence of histone H4 has remained almost invariant in more than 2 billion years of evolution [1,E1]. The region used as a signature pattern is a pentapeptide found in positions 14 to 18 of all H4sequences. It contains a lysine residue which is often acetylated [2] and a histidine residue which is implicated in DNA-binding [3].

Consensus pattern: G-A-K-R-H-

- [1] Thatcher T.H., Gorovsky M.A. Nucleic Acids Res. 22:174-179(1994).
[2] Doenecke D., Gallwitz D. Mol. Cell. Biochem. 44:113-128(1982).
[3] Ebralidse K.K., Grachev S.A., Mirzabekov A.D. Nature 331:365-367(1988).

Histone H3 signatures (his3)

- Histone H3 is one of the four histones, along with H2A, H2B and H4, which forms the eukaryotic nucleosome core. It is a highly conserved protein of 135 amino acid residues [1,2,E1]. The following proteins have been found to contain a C-terminal H3-like domain: - Mammalian centromeric protein CENP-A [3]. Could act as a core histone necessary for the assembly of centromeres. - Yeast chromatin-associated protein CSE4 [4]. - Caenorhabditis elegans chromosome III encodes two highly related proteins (F54C8.2 and F58A4.3) whose C-terminal section is evolutionary related to the last 100 residues of H3. The function of these proteins is not yet known. Two signature patterns were developed, The first one corresponds to a perfectly conserved heptapeptide in the N-terminal part of H3. The second one is derived from a conserved region in the central section of H3.

Consensus pattern: K-A-P-R-K-Q-L-

Consensus pattern: P-F-x-[RA]-L-[VA]-[KRQ]-[DEG]-[IV]-

- [1] Wells D.E., Brown D. Nucleic Acids Res. 19:2173-2188(1991).
[2] Thatcher T.H., Gorovsky M.A. Nucleic Acids Res. 22:174-179(1994).
[3] Sullivan K.F., Hechenberger M., Masri K. J. Cell Biol. 127:581-592(1994).
[4] Stoler S., Keith K.C., Curnick K.E., Fitzgerald-Hayes M. Genes Dev. 9:573-586(1995).

Histone H2B signature (his4)

Histone H2B is one of the four histones, along with H2A, H3 and H4, which forms the eukaryotic nucleosome core. Using alignments of histone H2B sequences [1,2,[E1](#)], a conserved region was selected in the C-terminal part of H2B.

5

Consensus pattern: [KR]-E-[LIVM SEQ ID NO:4)]-[EQ]-T-x(2)-[KR]-x-[LIVM SEQ ID NO:4)](2)-x-[PAG]-[DE]-L- x-[KR]-H-A-[LIVM SEQ ID NO:4)]-[STA]-E-G-

[1] Wells D.E., Brown D. Nucleic Acids Res. 19:2173-2188(1991).

10 [2] Thatcher T.H., Gorovsky M.A. Nucleic Acids Res. 22:174-179(1994).

297. 'Homeobox' domain signature and profile (home1)

The 'homeobox' is a protein domain of 60 amino acids [1 to 5,E1] first identified in a number of *Drosophila* homeotic and segmentation proteins. It has since been found to be extremely well conserved in many other animals, including vertebrates. This domain binds DNA through a helix-turn-helix type of structure. Some of the proteins which contain a homeobox domain play an important role in development. Most of these proteins are known to be sequence specific DNA-binding transcription factors. The homeobox domain has also been found to be very similar to a region of the yeast mating type proteins. These are sequence-specific DNA-binding proteins that act as master switches in yeast differentiation by controlling gene expression in a cell type-specific fashion. A schematic representation of the homeobox domain is shown below. The helix-turn-helix region is shown by the symbols 'H' (for helix), and 't' (for turn).

[illegible]

10 20 30 40 50 60 The pattern to detect homeobox sequences that was developed is 24 residues long and spans positions 34 to 57 of the homeobox domain.

30 Consensus pattern: [LIVMFY G SEQ ID NO:168)]-[ASLVR SEQ ID NO:321)]-x(2)-
[LIVMSTACN SEQ ID NO:322)]-x-[LIVM SEQ ID NO:4)]-x(4)-[LIV]- [RKNQESTAIY
SEQ ID NO:323)]-[LIVFSTNKH SEQ ID NO:324)]-W-[FYVC SEQ ID NO:239)]-x-
[NDQTAH SEQ ID NO:325)]-x(5)- [RKNAIMW SEQ ID NO:326)] -

[1] Gehring W.J. (In) Guidebook to the homeobox genes, Duboule D., Ed., pp1-10, Oxford University Press, Oxford, (1994).

[2] Buerklin T.R. (In) Guidebook to the homeobox genes, Duboule D., Ed., pp25-72, Oxford University Press, Oxford, (1994).

5 [3] Gehring W.J. Trends Biochem. Sci. 17:277-280(1992).

[4] Gehring W.J., Hiromi Y. Annu. Rev. Genet. 20:147-173(1986).

[5] Schofield P.N. Trends Neurosci. 10:3-6(1987).

'Homeobox' antennapedia-type protein signature (home2)

10 The homeotic Hox proteins are sequence-specific transcription factors. They are part of a developmental regulatory system that provides cells with specific positional identities on the anterior-posterior (A-P) axis [1]. The hox proteins contain a 'homeobox' domain. In *Drosophila* and other insects, there are eight different Hox genes that are encoded in two gene complexes, ANT-C and BX-C. In vertebrates there are 38 genes organized in four complexes.

15 In six of the eight *Drosophila* Hox genes the homeobox domain is highly similar and a conserved hexapeptide is found five to sixteen amino acids upstream of the homeobox domain. The six *Drosophila* proteins that belong to this group are antennapedia (Antp), abdominal-A (abd-A), deformed (Dfd), proboscipedia (pb), sex combs reduced (scr) and ultrabithorax (ubx) and are collectively known as the 'antennapedia' subfamily. In vertebrates

20 the corresponding Hox genes are known [2] as Hox-A2, A3, A4, A5, A6, A7, Hox-B1, B2, B3, B4, B5, B6, B7, B8, Hox-C4, C5, C6, C8, Hox-D1, D3, D4 and D8. *Caenorhabditis elegans* lin-39 and mab-5 are also members of the 'antennapedia' subfamily. As a signature pattern for this subfamily of homeobox proteins, the conserved hexapeptide was used.

25 Consensus pattern: [LIVMFE SEQ ID NO:327)]-[FY]-P-W-M-[KRQTA SEQ ID NO:328)]-

[1] McGinnis W., Krumlauf R. Cell 68:283-302(1992).

[2] Scott M.P. Cell 71:551-553(1992).

30 'Homeobox' engrailed-type protein signature (home3)

Most proteins which contain a 'homeobox' domain can be classified [1,2], on the basis of their sequence characteristics, in three subfamilies: engrailed, antennapedia and paired. Proteins currently known to belong to the engrailed subfamily are: - *Drosophila* segmentation

polarity protein engrailed (en) which specifies the body segmentation pattern and is required for the development of the central nervous system. - *Drosophila* invected protein (inv). - Silk moth proteins engrailed and invected, which may be involved in the compartmentalization of the silk gland. - Honeybee E30 and E60. - Grasshopper (*Schistocerca americana*) G-En. -
5 Mammalian and birds En-1 and En-2. - Zebrafish Eng-1, -2 and -3. - Sea urchin (*Tripneustes gratilla*) SU-HB-en. - Leech (*Helobdella triserialis*) Ht-En. - *Caenorhabditis elegans* ceh-
16. Engrailed homeobox proteins are characterized by the presence of a conserved region of some 20 amino-acid residues located at the C-terminal of the 'homeobox' domain. As a signature pattern for this subfamily of proteins, a stretch of eight perfectly conserved residues
10 in this region was used.

Consensus pattern: L-M-A-[EQ]-G-L-Y-N-

[1] Scott M.P., Tamkun J.W., Hartzell G.W. III *Biochim. Biophys. Acta* 989:25-48(1989).

15 [2] Gehring W.J. *Science* 236:1245-1252(1987).

298. Isocitrate lyase signature (ICL)

Isocitrate lyase (EC 4.1.3.1) [1,2] is an enzyme that catalyzes the conversion of isocitrate to
20 succinate and glyoxylate. This is the first step in the glyoxylate bypass, an alternative to the tricarboxylic acid cycle in bacteria, fungi and plants. A cysteine, a histidine and a glutamate or aspartate have been found to be important for the enzyme's catalytic activity. Only one cysteine residue is conserved between the sequences of the fungal, plant and bacterial enzymes; it is located in the middle of a conserved hexapeptide that can be used as a
25 signature pattern for this type of enzyme.

Consensus pattern: K-[KR]-C-G-H-[LMQ] [C is a putative active site residue]-

[1] Beeching J.R. *Protein Seq. Data Anal.* 2:463-466(1989).

30 [2] Atomi H., Ueda M., Hikida M., Hishida T., Teranishi Y., Tanaka A. *J. Biochem.* 107:262-266(1990).

299. Initiation factor 2 subunit

This family includes initiation factor 2B alpha, beta and delta subunits from eukaryotes, related proteins from archaeobacteria and IF-2 from prokaryotes. Initiation factor 2 binds to Met-tRNA, GTP and the small ribosomal subunit.

- 5 [1] Kyripides NC, Woese CR, Proc Natl Acad Sci U S A 1998;95:3726-3730.

300. Initiation factor 3 signature

10 Initiation factor 3 (IF-3) (gene infC) [1] is one of the three factors required for the initiation of protein biosynthesis in bacteria. IF-3 is thought to function as a fidelity factor during the assembly of the ternary initiation complex which consist of the 30S ribosomal subunit, the initiator tRNA and the messenger RNA. IF-3 binds to the 30S ribosomal subunit; it is a basic protein of 141 to 212 residues. The chloroplast initiation factor IF-3(chl) is a protein that enhances the poly(A,U,G)-dependent binding of the initiator tRNA to chloroplast

15 ribosomal30s subunits. In its mature form it is a protein of about 400 residues whose central section is evolutionary related to the sequence of bacterial IF-3 [2].As a signature pattern a highly conserved region was selected located in the central section of bacterial IF-3 and of IF-3(chl).

20 Consensus pattern: [KR]-[LIVM SEQ ID NO:4)](2)-[DN]-[FY]-[GSN]-[KR]-[LIVMFYS SEQ ID NO:153)]-x-[FY]- [DEQTH SEQ ID NO:329)]-x(2)-[KRQ]-

[1] Liveris D., Schwartz J.J., Geertman R., Schwartz I. FEMS Microbiol. Lett. 112:211-216(1993).

25 [2] Lin Q., Ma L., Burkhardt W., Spremulli L.L. J. Biol. Chem. 269:9436-9444(1994).

301. Imidazoleglycerol-phosphate dehydratase signatures (IGPD)

30 Imidazoleglycerol-phosphate dehydratase (EC 4.2.1.19) is the enzyme that catalyzes the seventh step in the biosynthesis of histidine in bacteria, fungi and plants. In most organisms it is a monofunctional protein of about 22 to 29 Kd. In some bacteria such as Escherichia coli it is the C-terminal domain of a bifunctional protein that include a histidinol-phosphatase

domain [1]. Two signature patterns were developed that each include two consecutive histidine residues.

Consensus pattern: [LIVMY SEQ ID NO:141)]-[DE]-x-H-H-x(2)-E-x(2)-[GCA]-[LIVM
5 SEQ ID NO:4)]-[STAC SEQ ID NO:204)]-[LIVM SEQ ID NO:4)]-
Consensus pattern: G-x-[DN]-x-H-H-x(2)-E-[STAGC SEQ ID NO:45)]-x-[FY]-K -

[1] Carlomagno M.S., Chiariotti L., Alifano P., Nappo A.G., Bruni C.B. J. Mol. Biol.
203:585-606(1988).

302. Indole-3-glycerol phosphate synthase signature (IGPS)

Indole-3-glycerol phosphate synthase (EC 4.1.1.48) (IGPS) catalyzes the fourth step in the
biosynthesis of tryptophan: the ring closure of 1-(2-carboxy-phenylamino)-1-deoxyribose
15 into indol-3-glycerol-phosphate. In some bacteria, IGPS is a single chain enzyme. In others -
such as *Escherichia coli* - it is the N-terminal domain of a bifunctional enzyme that also
catalyzes N-(5'-phosphoribosyl)anthranilate isomerase (PRAI) activity, the third step of
tryptophan biosynthesis. In fungi, IGPS is the central domain of a trifunctional enzyme that
also contains a PRAI C-terminal domain and a glutamine amidotransferase N-terminal
20 domain. The N-terminal section of IGPS contains a highly conserved region which X-ray
crystallography studies [1] have shown to be part of the active site cavity. This region was
used as a signature pattern for IGPS.

Consensus pattern: [LIVMFY SEQ ID NO:18)]-[LIVMC SEQ ID NO:142)]-x-E-[LIVMFYC
25 SEQ ID NO:6)]-K-[KRSP SEQ ID NO:330)]-[STAK SEQ ID NO:331)]-S-P-[ST]- x(3)-
[LIVMFYST SEQ ID NO:332)]-

[1] Wilmanns M., Priestle J.P., Niemann T., Jansonius J.N. J. Mol. Biol. 223:477-
507(1992).

303. (IL2) Interleukin 2. 31 members

304. (ILVD EDD) Dihydroxy-acid and 6-phosphogluconate dehydratases. Two dehydratases have been shown [1] to be evolutionary related: - Dihydroxy-acid dehydratase (EC 4.2.1.9) (gene *ilvD* or *ILV3*) which catalyzes the fourth step in the biosynthesis of isoleucine and valine, the dehydration of 2,3-dihydroxy-isovaleric acid into alpha-ketoisovaleric acid. - 6-phosphogluconate dehydratase (EC 4.2.1.12) (gene *edd*) which catalyzes the first step in the Entner-Doudoroff pathway, the dehydration of 6-phospho- D-gluconate into 6-phospho-2-dehydro-3-deoxy-D-gluconate. - *Escherichia coli* hypothetical protein *yjhG*. Both enzymes are proteins of about 600 amino acid residues. Two highly conserved regions have been developed as signature patterns. The first pattern is located in the N-terminal part and contains a cysteine that could be involved in the binding of a 2Fe-2S iron-sulfur cluster [2]. The second pattern is located in the C-terminal half.

Consensus pattern: C-D-K-x(2)-P-[GA]-x(3)-[GA] [The C could be a 2Fe-2S ligand]

Consensus pattern: [SA]-L-[LIVM SEQ ID NO:4)]-T-D-[GA]-R-[LIVMF SEQ ID NO:2)]-S-[GA]-[GAV]-[ST]-

[1] Egan S.E., Fliege R., Tong S., Shibata A., Wolf R.E. Jr., Conway T. J. *Bacteriol.*

174:4638-4646(1992).[2] Velasco J.A., Cansado J., Pena M.C., Kawakami T., Laborda J.,

Notario V. *Gene* 137:179-185(1993).

305. IMP dehydrogenase / GMP reductase signature

IMP dehydrogenase (EC 1.1.1.205) (IMPDH) catalyzes the rate-limiting reaction of de novo GTP biosynthesis, the NAD-dependent reduction of IMP into XMP [1]. Inhibition of IMP dehydrogenase activity results in the cessation of DNA synthesis. As IMP dehydrogenase is associated with cell proliferation, it is a possible target for cancer chemotherapy. Mammalian and bacterial IMPDHs are tetramers of identical chains. There are two IMP dehydrogenase isozymes in humans [2]. GMP reductase (EC 1.6.6.8) catalyzes the irreversible and NADPH-dependent reductive deamination of GMP into IMP [3]. It converts nucleobase, nucleoside and nucleotide derivatives of G to A nucleotides, and maintains intracellular balance of A and G nucleotides. IMP dehydrogenase and GMP reductase share many regions of sequence

similarity. One of these regions is centered on a cysteine residue thought [3] to be involved in binding IMP. This region was used as a signature pattern.

Consensus pattern: [LIVM SEQ ID NO:4)]-[RK)]-[LIVM SEQ ID NO:4)]-G-[LIVM SEQ ID NO:4)]-G-x-G-S-[LIVM SEQ ID NO:4)]-C-x-T [C is the putative IMP-binding residue]-

[1] Collart F.R., Huberman E. J. Biol. Chem. 263:15769-15772(1988).

[2] Natsumeda Y., Ohno S., Kawasaki H., Konno Y., Weber G., Suzuki K. J. Biol. Chem. 265:5292-5295(1990).

[3] Andrews S.C., Guest J.R. Biochem. J. 255:35-43(1988).

306. (IPPC) Inositol polyphosphate phosphatase family, catalytic domain

[1] York JD, Ponder JW, Chen ZW, Mathews FS, Majerus PW; Biochemistry 1994;33:13164-13171. [2] Jefferson AB, Auethavekiat V, Pot DA, Williams LT, Majerus PW; J Biol Chem 1997;272:5983-5988. [3] Zhang X, Jefferson AB, Auethavekiat V, Majerus PW; Proc Natl Acad Sci U S A 1995;92:4853-4856. [4] York JD, Majerus PW. Proc Natl Acad Sci U S A 1990;87:9548-9552. [5] Neuwald AF, York JD, Majerus PW; FEBS Lett 1991;294:16-18.

307. IQ calmodulin-binding motif

[1] Xie X, Harrison DH, Schlichting I, Sweet RM, Kalabokis VN, Szent-Gyorgyi AG, Cohen C; Nature 1994;368:306-312.
[2] Rhoads AR, Friedberg F; FASEB J 1997;11:331-340.

308. Inosine-uridine preferring nucleoside hydrolase family signature (IU nuc hydro)
Inosine-uridine preferring nucleoside hydrolase (EC 3.2.2.1) (IU-nucleosidehydrolase or IUNH) is an enzyme first identified in protozoan [1] that catalyzes the hydrolysis of all of the

commonly occurring purine and pyrimidine nucleosides into ribose and the associated base, but has a preference for inosine and uridine as substrates. This enzyme is important for these parasitic organisms, which are deficient in de novo synthesis of purines, to salvage the host purine nucleosides. IUNH from *Crithidia fasciculata* has been sequenced and characterized, it is an homotetrameric enzyme of subunits of 34 Kd. An histidine has been shown to be important for the catalytic mechanism, it acts a proton donor to activate the hypoxanthine leaving group. IUNH is evolutionary related to a number of uncharacterized proteins from various biological sources, notably: - *Escherichia coli* hypothetical protein yaaF. - *Escherichia coli* hypothetical protein ybeK. - *Escherichia coli* hypothetical protein yeiK. - Fission yeast hypothetical protein SpAC17G8.02. - Yeast hypothetical protein YDR400w. - An hypothetical protein from the archaebacteria *Desulfurolobus ambivalens*. As a signature pattern for these proteins, a highly conserved region was selected located in the N-terminal extremity. This region contains four conserved aspartates that have been shown [2] to be located in the active site cavity.

Consensus pattern: D-x-D-[PT]-[GA]-x-D-D-[TAV]-[VI]-A -

[1] Gopaul D.N., Meyer S.L., Degano M., Sacchettini J.C., Schramm V.L. *Biochemistry* 35:5963-5970(1996).

[2] Degano M., Gopaul D.N., Scapin G., Schramm V.L., Sacchettini J.C. *Biochemistry* 35:5971-5981(1996).

309. (Insulinase)

Insulinase family, zinc-binding region signature
(aka Peptidase_M16)

A number of proteases dependent on divalent cations for their activity have been shown [1,2] to belong to one family, on the basis of sequence similarity. These enzymes are listed below.

- Insulinase (EC 3.4.24.56) (also known as insulysin or insulin-degrading enzyme or IDE), a cytoplasmic enzyme which seems to be involved in the cellular processing of insulin, glucagon and other small polypeptides.

- *Escherichia coli* protease III (EC 3.4.24.55) (pitrylsin) (gene ptr), a periplasmic enzyme that degrades small peptides.

- Mitochondrial processing peptidase (EC 3.4.24.64) (MPP). This enzyme removes the transit peptide from the precursor form of proteins imported from the cytoplasm across the mitochondrial inner membrane. It is composed of two nonidentical homologous subunits termed alpha and beta. The beta subunit seems to be catalytically active while the alpha subunit has probably lost its activity.

- Nardilysin (EC 3.4.24.61) (N-arginine dibasic convertase or NRD convertase) this mammalian enzyme cleaves peptide substrates on the N-terminus of Arg residues in dibasic stretches.

- *Klebsiella pneumoniae* protein pqqF. This protein is required for the biosynthesis of the coenzyme pyrrolo-quinoline-quinone (PQQ). It is thought to be protease that cleaves peptide bonds in a small peptide (gene pqqA) thus providing the glutamate and tyrosine residues necessary for the synthesis of PQQ.

- Yeast protein AXL1, which is involved in axial budding [3].

- *Eimeria bovis* sporozoite developmental protein.

- *Escherichia coli* hypothetical protein yddC and HI1368, the corresponding *Haemophilus influenzae* protein.

- *Bacillus subtilis* hypothetical protein ymxG.

- *Caenorhabditis elegans* hypothetical proteins C28F5.4 and F56D2.1.

It should be noted that in addition to the above enzymes, this family also includes the core proteins I and II of the mitochondrial bc1 complex (also called cytochrome c reductase or complex III), but the situation as to the activity or lack of activity of these subunits is quite complex:

- In mammals and yeast, core proteins I and II lack enzymatic activity.

- In *Neurospora crassa* and in potato core protein I is equivalent to the beta subunit of MPP.

- In *Euglena gracilis*, core protein I seems to be active, while subunit II is inactive.

These proteins do not share many regions of sequence similarity; the most noticeable is in the N-terminal section. This region includes a conserved histidine followed, two residues later by a glutamate and another histidine. In pitrylsin, it has been shown [4] that this H-x-x-E-H

motif is involved in enzyme activity; the two histidines bind zinc and the glutamate is necessary for catalytic activity. Non active members of this family have lost from one to three of these active site residues. We developed a signature pattern that detect active members of this family as well as some inactive members.

5

Consensus pattern G-x(8,9)-G-x-[STA]-H-[LIVMFY SEQ ID NO:18)]-[LIVMC SEQ ID NO:142)]-[DERN SEQ ID NO:333)]-[HRKL SEQ ID NO:334)]- [LMFAT SEQ ID NO:335)]-x-[LFSTH SEQ ID NO:336)]-x-[GSTAN SEQ ID NO:296)]-[GST] [The two H are zinc ligands] [E is the active site residue] Sequences known to belong to this class detected by the pattern ALL active members as well as all MPP alpha subunits and core II subunits. Does not detect inactive core I subunits.

10

Note: these proteins belong to family M16 in the classification of peptidases [5].

15

[1] Rawlings N.D., Barrett A.J. Biochem. J. 275:389-391(1991).

[2] Braun H.-P., Schmitz U.K. Trends Biochem. Sci. 20:171-175(1995).

[3] Becker A.B., Roth R.A. Proc. Natl. Acad. Sci. U.S.A. 89:3835-3839(1992).

20

[4] Fujita A., Oka C., Arikawa Y., Katagai T., Tonouchi A., Kuhara S., Misumi Y. Nature 372:567-570(1994).

[5] Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-228(1995).

25

310. Involucrin repeat

Eckert RL, Yaffe MB, Crish JF, Murthy S, Rorke EA, Welter JF, J Invest Dermatol 1993;100:613-617.

30

311. Isochorismatase family. This family are hydrolase enzymes.

Romao MJ, Turk D, Gomis-Ruth FX, Huber R, Schumacher G, Mollering H, Russmann L, J Mol Biol 1992;226:1111-1130.

312. Inositol monophosphatase family signatures (inositol_P)

It has been shown [1] that several proteins share two sequence motifs. Two of these proteins
 5 are enzymes of the inositol phosphate second messenger signaling pathway: - Vertebrate and
 plants inositol monophosphatase (EC 3.1.3.25). - Vertebrate inositol polyphosphate 1-
 phosphatase (EC 3.1.3.57). The function of the other proteins is not yet clear: - Bacterial
 protein *cysQ*. *CysQ* could help to control the pool of PAPS (3'-phosphoadenoside 5'-
 phosphosulfate), or be useful in sulfite synthesis. - *Escherichia coli* protein *suhB*. Mutations
 10 in *suhB* results in the enhanced synthesis of heat shock sigma factor (*htrR*). - *Neurospora*
crassa protein *Qa-X*. Probably involved in quinate metabolism. - *Emericella nidulans* protein
qutG. Probably involved in quinate metabolism. - Yeast protein *HAL2/MET22* [2] involved
 in salt tolerance as well as methionine biosynthesis. - Yeast hypothetical protein
YHR046c. - *Caenorhabditis elegans* hypothetical protein *F13G3.5*. - A *Rhizobium*
 15 *leguminosarum* hypothetical protein encoded upstream of the *pss* gene for exopolysaccharide
 synthesis. - *Methanococcus jannaschii* hypothetical protein *MJ0109*. It is suggested [1] that
 these proteins may act by enhancing the synthesis or degradation of phosphorylated
 messenger molecules. From the X-ray structure of human inositol monophosphatase [3], it
 seems that some of the conserved residues are involved in binding a metal ion and/or the
 20 phosphate group of the substrate.

Consensus pattern: [FWV]-x(0,1)-[LIVM SEQ ID NO:4]-D-P-[LIVM SEQ ID NO:4]-D-
 [SG]-[ST]-x(2)-[FY]-x- [HKRNSTY SEQ ID NO:337)] [The first D and the T bind a metal
 ion]-

25 Consensus pattern: [WV]-D-x-[AC]-[GSA]-[GSAPV SEQ ID NO:338)]-x-[LIVACP SEQ ID
 NO:339)]-[LIV]-[LIVAC SEQ ID NO:319)]-x(3)- [GH]-[GA]-

[1] Neuwald A.F., York J.D., Majerus P.W. FEBS Lett. 294:16-18(1991).

[2] Glaeser H.-U., Thomas D., Gaxiola R., Montrichard F., Surdin-Kerjan Y., Serrano R.
 30 EMBO J. 12:3105-3110(1993).

[3] Bone R., Springer J.P., Atack J.R. Proc. Natl. Acad. Sci. U.S.A. 89:10031-10035(1992).

313. Ion transport protein

This family contains Sodium, Potassium, Calcium ion channel. This family is 6 transmembrane helices in which the last two helices flank a loop which determines ion selectivity. In some sub-families (e.g. Na channels) the domain is repeated four times, whereas in others (e.g. K channels) the protein forms as a tetramer in the membrane. A bacterial structure of the protein is known for the last two helices but is not the Pfam family due to it lacking the first four helices.

314. Isocitrate and isopropylmalate dehydrogenases signature (isodh)

Isocitrate dehydrogenase (IDH) [1,2] is an important enzyme of carbohydrate metabolism which catalyzes the oxidative decarboxylation of isocitrate into alpha-ketoglutarate. IDH is either dependent on NAD⁺ (EC 1.1.1.41) or on NADP⁺ (EC 1.1.1.42). In eukaryotes there are at least three isozymes of IDH: two are located in the mitochondrial matrix (one NAD⁺-dependent, the other NADP⁺-dependent), while the third one (also NADP⁺-dependent) is cytoplasmic. In *Escherichia coli* the activity of a NADP⁺-dependent form of the enzyme is controlled by the phosphorylation of a serine residue; the phosphorylated form of IDH is completely inactivated. 3-isopropylmalate dehydrogenase (EC 1.1.1.85) (IMDH) [3,4] catalyzes the third step in the biosynthesis of leucine in bacteria and fungi, the oxidative decarboxylation of 3-isopropylmalate into 2-oxo-4-methylvalerate. Tartrate dehydrogenase (EC 1.1.1.93) [5] catalyzes the reduction of tartrate to oxaloglycolate. These enzymes are evolutionary related [1,3,4,5]. The best conserved region of these enzymes is a glycine-rich stretch of residues located in the C-terminal section. This region was used as a signature pattern.

Consensus pattern: [NS]-[LIMYT SEQ ID NO:340)]-[FYDN SEQ ID NO:341)]-G-[DNT]-[IMVY SEQ ID NO:342)]-x-[STGDN SEQ ID NO:206)]-[DN]-x(2)-[SGAP SEQ ID NO:343)]-x(3,4)-G-[STG]-[LIVMPA SEQ ID NO:344)]-G-[LIVMF SEQ ID NO:2)]-

[1] Hurley J.H., Thorsness P.E., Ramalingam V., Helmers N.H., Koshland D.E. Jr., Stroud R.M. *Proc. Natl. Acad. Sci. U.S.A.* 86:8635-8639(1989).

[2] Cupp J.R., McAlister-Henn L. *J. Biol. Chem.* 266:22199-22205(1991).

315

[3] Imada K., Sato M., Tanaka N., Katsube Y., Matsuura Y., Oshima T. J. Mol. Biol. 222:725-738(1991).

[4] Zhang T., Koshland D.E. Jr. Protein Sci. 4:84-92(1995).

[5] Tipton P.A., Beecher B.S. Arch. Biochem. Biophys. 313:15-21(1994).

5

315. Jacalin-like lectin domain.

Proteins containing this domain are lectins. It is found in

10 1 to 6 copies in these proteins. The domain is also found in the animal prostatic spermine-binding protein (Swiss:P15501).

[1] Sankaranarayanan R, Sekar K, Banerjee R, Sharma V, Surolia A, Vijayan M; Nat Struct Biol 1996;3:596-603.

15

316. KH domain

KH motifs probably bind RNA directly. Auto antibodies to Nova, a KH domain protein, cause paraneoplastic opsoclonus ataxia.

20 [1] Burd CG, Dreyfuss G, Science 1994;265:615-621.

[2] Musco G, Stier G, Joseph C, Castiglione Morelli MA, Nilges M, Gibson TJ, Pastore A, Cell 1996;85:237-245.

25 317. Kelch motif

The kelch motif was initially discovered in Kelch (Swiss:Q04652). In this protein there are six copies of the motif. It has been shown that Swiss:Q04652 is related to Galactose Oxidase [1] for which a structure has been solved [2]. The kelch motif forms a beta sheet. Several of these sheets associate to form a beta propeller structure as found in neur,

30 [1] Bork P, Doolittle RF, J Mol Biol 1994;236:1277-1282. [2] Ito N, Phillips SE, Stevens C, Ogel ZB, McPherson MJ, Keen, JN, Yadav KD, Knowles PF, Nature 1991;350:87-90.

318. Soybean trypsin inhibitor (Kunitz) protease inhibitors family signature

The soybean trypsin inhibitor (Kunitz) family [1] is one of the numerous families of proteinase inhibitors. It comprise plant proteins which have inhibitory activity against serine proteinases from the trypsin and subtilisin families, thiol proteinases and aspartic proteinases as well as some proteins that are probably involved in seed storage. This family is currently known to group the following proteins: - Trypsin inhibitors A, B, C, KTI1, and KTI2 from soybean. - Trypsin inhibitor DE3 from coral beans (*Erythrina* sp.). - Trypsin inhibitor DE5 from sandal bead tree. - Trypsin inhibitors 1A (WTI-1A), 1B (WTI-1B), and 2 (WTI-2) from goa bean. - Trypsin inhibitor from *Acacia confusa*. - Trypsin inhibitor from silk tree. - Chymotrypsin inhibitor 3 (WCI-3) from goa bean. - Cathepsin D inhibitors PDI and NDI from potato [2], which inhibit both cathepsin D (aspartic proteinase) and trypsin. - Alpha-amylase/subtilisin inhibitors from barley and wheat. - Albumin-1 (WBA-1) from goa bean seeds [3]. - Miraculin from *Richadella dulcifica* [4], a sweet taste protein. - Sporamin from sweet potato [5], the major tuberous root protein. - Thiol proteinase inhibitor PCPI 8.3 (P340) from potato tuber [6]. - Wound responsive protein gwin3 from poplar tree [7]. - 21 Kd seed protein from cocoa [8]. All these proteins contain from 170 to 200 amino acid residues and one or two intrachain disulfide bonds. The best conserved region is found in their N-terminal section and is used as a signature pattern

Consensus pattern: [LIVM SEQ ID NO:4)]-x-D-x-[EDNTY SEQ ID NO:345)]-[DG]-[RKHDENQ SEQ ID NO:346)]-x-[LIVM SEQ ID NO:4)]-x(5)-Y-x-[LIVM SEQ ID NO:4)]

-

[1] Laskowski M., Kato I. *Annu. Rev. Biochem.* 49:593-626(1980).

[2] Ritonja A., Krizaj I., Mesko P., Kopitar M., Lucovnik P., Strukelj B., Pungercar J., Buttle D.J., Barrett A.J., Turk V. *FEBS Lett.* 267:13-15(1990).

[3] Kortt A.A., Strike P.M., de Jersey J. *Eur. J. Biochem.* 181:403-408(1989).

[4] Theerasilp S., Hitotsuya H., Nakajo S., Nakaja K., Nakamura Y., Kurihara Y. *J. Biol.*

Chem. 264:6655-6659(1989).

[5] Hattori T., Yoshida N., Nakamura K. *Plant Mol. Biol.* 13:563-572(1989).

[6] Krizaj I., Drobic-Kosorok M., Brzin J., Jerala R., Turk V. *FEBS Lett.* 333:15-20(1993).

[7] Bradshaw H.D., Hollick J.B., Parsons T.J., Clarke H.R.G., Gordon M.P. Plant Mol. Biol. 14:51-59(1989).

[8] Tai H., McHenry L., Fritz P.J., Furtek D.B. Plant Mol. Biol. 16:913-915(1991).

5

319. Beta-ketoacyl synthases active site

Beta-ketoacyl-ACP synthase (KAS) [1] is the enzyme that catalyzes the condensation of malonyl-ACP with the growing fatty acid chain. It is found as a component of the following enzymatic systems: - Fatty acid synthetase (FAS), which catalyzes the formation of long-chain fatty acids from acetyl-CoA, malonyl-CoA and NADPH. Bacterial and plant chloroplast FAS are composed of eight separate subunits which correspond to different enzymatic activities; beta-ketoacyl synthase is one of these polypeptides. Fungal FAS consists of two multifunctional proteins, FAS1 and FAS2; the beta-ketoacyl synthase domain is located in the C-terminal section of FAS2. Vertebrate FAS consists of a single multifunctional chain; the beta-ketoacyl synthase domain is located in the N-terminal section [2]. - The multifunctional 6-methylsalicylic acid synthase (MSAS) from *Penicillium patulum* [3]. This is a multifunctional enzyme involved in the biosynthesis of a polyketide antibiotic and which has a KAS domain in its N-terminal section. - Polyketide antibiotic synthase enzyme systems. Polyketides are secondary metabolites produced by microorganisms and plants from simple fatty acids. KAS is one of the components involved in the biosynthesis of the *Streptomyces* polyketide antibiotics granatacin [4], tetracenomycin C [5] and erythromycin. - *Emericella nidulans* multifunctional protein Wa. Wa is involved in the biosynthesis of conidial green pigment. Wa is protein of 216 Kd that contains a KAS domain. - *Rhizobium* nodulation protein nodE, which probably acts as a beta-ketoacyl synthase in the synthesis of the nodulation Nod factor fatty acyl chain. - Yeast mitochondrial protein CEM1. The condensation reaction is a two step process: the acyl component of an activated acyl primer is transferred to a cysteine residue of the enzyme and is then condensed with an activated malonyl donor with the concomitant release of carbon dioxide. The sequence around the active site cysteine is well conserved and can be used as a signature pattern.

30

Consensus pattern: G-x(4)-[LIVMFAP SEQ ID NO:347])-x(2)-[AGC]-C-[STA](2)-[STAG SEQ ID NO:20)]-x(3)-[LIVMF SEQ ID NO:2)] [C is the active site residue]

[1] Kauppinen S., Siggaard-Andersen M., von Wettstein-Knowles P. Carlsberg Res. Commun. 53:357-370(1988).

[2] Witkowski A., Rangan V.S., Randhawa Z.I., Amy C.M., Smith S. Eur. J. Biochem. 198:571-579(1991).

5 [3] Beck J., Ripka S., Siegner A., Schiltz E., Schweizer E. Eur. J. Biochem. 192:487-498(1990).

[4] Bibb M.J., Biro S., Motamedi H., Collins J.F., Hutchinson C.R. EMBO J. 8:2727-2736(1989).

10 [5] Sherman D.H., Malpartida F., Bibb M.J., Kieser H.M., Bibb M.J., Hopwood D.A. EMBO J. 8:2717-2725(1989).

320. Kinesin motor domain signature and profile

Kinesin [1,2,3] is a microtubule-associated force-producing protein that may play a role in organelle transport. Kinesin is an oligomeric complex composed of two heavy chains and two light chains. The kinesin motor activity is directed toward the microtubule's plus end. The heavy chain is composed of three structural domains: a large globular N-terminal domain which is responsible for the motor activity of kinesin (it is known to hydrolyze ATP, to bind and move on microtubules), a central alpha-helical coiled coil domain that mediates the heavy chain dimerization; and a small globular C-terminal domain which interacts with other proteins (such as the kinesin light chains), vesicles and membranous organelles. A number of proteins have been recently found that contain a domain similar to that of the kinesin 'motor' domain [1,4,E1]: - *Drosophila* claret segregational protein (ncd). Ncd is required for normal chromosomal segregation in meiosis, in females, and in early mitotic divisions of the embryo. The ncd motor activity is directed toward the microtubule's minus end. - *Drosophila* kinesin-like protein (nod). Nod is required for the distributive chromosome segregation of nonexchange chromosomes during meiosis. - Human CENP-E [4]. CENP-E is a protein that associates with kinetochores during chromosome congression, relocates to the spindle midzone at anaphase, and is quantitatively discarded at the end of the cell division. CENP-E is probably an important motor molecule in chromosome movement and/ or spindle elongation. - Human mitotic kinesin-like protein-1 (MKLP-1), a motor protein whose activity is directed toward the microtubule's plus end. - Yeast KAR3 protein, which is essential for yeast nuclear fusion during mating. KAR3 may mediate microtubule sliding during nuclear

fusion and possibly mitosis. - Yeast CIN8 and KIP1 proteins which are required for the assembly of the mitotic spindle. Both proteins seem to interact with spindle microtubules to produce an outwardly directed force acting upon the poles. - Fission yeast cut7 protein, which is essential for spindle body duplication during mitotic division. - *Emericella nidulans* bimC, which plays an important role in nuclear division. - *Emericella nidulans* klpA. - *Caenorhabditis elegans* unc-104, which may be required for the transport of substances needed for neuronal cell differentiation. - *Caenorhabditis elegans* osm-3. - *Xenopus* Eg5, which may be involved in mitosis. - *Arabidopsis thaliana* KatA, KatB and katC. - *Chlamydomonas reinhardtii* FLA10/KHP1 and KLP1. Both proteins seem to play a role in the rotation or twisting of the microtubules of the flagella. - *Caenorhabditis elegans* hypothetical protein T09A5.2. The kinesin motor domain is located in the N-terminal part of most of the above proteins, with the exception of KAR3, klpA, and ncd where it is located in the C-terminal section. The kinesin motor domain contains about 330 amino acids. An ATP-binding motif of type A is found near position 80 to 90, the C-terminal half of the domain is involved in microtubule-binding. The signature pattern for that domain is derived from a conserved decapeptide inside the microtubule-binding part.

Consensus pattern: [GSA]-[KRHPSTQVM SEQ ID NO:348)]-[LIVMF SEQ ID NO:2)]-x-[LIVMF SEQ ID NO:2)]-[IVC]-D-L-[AH]-G-[SAN]-E

- [1] Bloom G.S., Endow S.A. Protein Prof. 2:1109-1171(1995).
- [2] Vallee R.B., Shpetner H.S. Annu. Rev. Biochem. 59:909-932(1990).
- [3] Brady S.T. Trends Cell Biol. 5:159-164(1995).
- [4] Endow S.A. Trends Biochem. Sci. 16:221-225(1991).[E1]

321. Ribosomal protein L15 signature

Ribosomal protein L15 is one of the proteins from the large ribosomal subunit. In *Escherichia coli*, L15 is known to bind the 23S rRNA. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups: - Eubacterial L15. - Plant chloroplast L15 (nuclear-encoded). - Archaeobacterial L15. - Vertebrate L27a. - *Tetrahymena thermophila* L29. - Fungi L27a (L29, CRP-1, CYH2). L15 is a protein of 144 to 154 amino-acid residues.

As a signature pattern, a conserved region was selected in the C-terminal section of these proteins.

Consensus pattern: K-[LIVM SEQ ID NO:4)](2)-[GASL SEQ ID NO:349)]-x-[GT]-x-
 5 [LIVMA SEQ ID NO:30)]-x(2,5)-[LIVM SEQ ID NO:4)]-x- [LIVMF SEQ ID NO:2)]-
 x(3,4)-[LIVMFCA SEQ ID NO:350)]-[ST]-x(2)-A-x(3)-[LIVM SEQ ID NO:4)]-x(3)-G

[1] Otake E., Hashimoto T., Mizuta K., Suzuki K. Protein Seq. Data Anal. 5:301-313(1993).

322. LBP / BPI / CETP family signature

The following mammalian lipid-binding serum glycoproteins belong to the same family [1,2,3]: - Lipopolysaccharide-binding protein (LBP). LBP binds to the lipid A moiety of bacterial lipopolysaccharides (LPS), a glycolipid present in the outer membrane of all Gram-
 15 negative bacteria. The LBP/LPS complex seems to interact with the CD14 receptor and may be responsible for the secretion of alpha-TNF. - Bactericidal permeability-increasing protein (BPI). Like LBP, BPI binds LPS and has a cytotoxic activity on Gram-negative bacteria. - Cholesteryl ester transfer protein (CETP). CETP is involved in the transfer of insoluble cholesteryl esters in reverse cholesterol transport. - Phospholipid transfer protein (PLTP).
 20 May play a key role in extracellular phospholipid transport and modulation of HDL particles. These proteins are structurally related and share many regions of sequence similarities. As a signature pattern one of these regions was selected, which is located in the N-terminal section of these proteins; a region which could be involved in the binding to the lipids [2].

25 Consensus pattern: [PA]-[GA]-[LIVMC SEQ ID NO:142)]-x(2)-R-[IV]-[ST]-x(3)-L-x(5)-
 [EQ]-x(4)- [LIVM SEQ ID NO:4)]-[EQK]-x(8)-P

[1] Schumann R.R., Leong S.R., Flaggs G.W., Gray P.W., Wright S.D., Mathison J.C., Tobias P.S., Ulevitch R.J. Science 249:1429-1431(1990).

30 [2] Gray P.W., Flaggs G., Leong S.R., Gumina R.J., Weiss J., Ooi C.E., Elsbach P. J. Biol. Chem. 264:9505-9509(1989).

[3] Day J.R., Albers J.J., Lofton-Day C.E., Gilbert T.L., Ching A.F.T., Grant F.J., O'Hara P.J., Marcovina S.M., Adolphson J.L. J. Biol. Chem. 269:9388-9391(1994).

323. LIM domain signature and profile

Recently [1,2] a number of proteins have been found to contain a conserved cysteine-rich domain of about 60 amino-acid residues. These proteins are: - *Caenorhabditis elegans* mec-3; a protein required for the differentiation of the set of six touch receptor neurons in this nematode. - *Caenorhabditis elegans* lin-11; a protein required for the asymmetric division of vulval blast cells. - Vertebrate insulin gene enhancer binding protein isl-1. Isl-1 binds to one of the two cis-acting protein-binding domains of the insulin gene. - Vertebrate homeobox proteins lim-1, lim-2 (lim-5) and lim3. - Vertebrate lmx-1, which acts as a transcriptional activator by binding to the FLAT element; a beta-cell-specific transcriptional enhancer found in the insulin gene. - Mammalian LH-2, a transcriptional regulatory protein involved in the control of cell differentiation in developing lymphoid and neural cell types. - *Drosophila* protein apterous, required for the normal development of the wing and halter imaginal discs. - Vertebrate protein kinases LIMK-1 and LIMK-2. - Mammalian rhombotins. Rhombotin 1 (RBTN1 or TTG-1) and rhombotin-2 (RBTN2 or TTG-2) are proteins of about 160 amino acids whose genes are disrupted by chromosomal translocations in T-cell leukemia. - Mammalian and avian cysteine-rich protein (CRP), a 192 amino-acid protein of unknown function. Seems to interact with zyxin. - Mammalian cysteine-rich intestinal protein (CRIP), a small protein which seems to have a role in zinc absorption and may function as an intracellular zinc transport protein. - Vertebrate paxillin, a cytoskeletal focal adhesion protein. - Mouse testin. Mouse testin should not be confused with rat testin which is a thiol protease homolog. - Sunflower pollen specific protein SF3. - Chicken zyxin. Zyxin is a low-abundance adhesion plaque protein which has been shown to interact with CRP. - Yeast protein LRG1 which is involved in sporulation [4]. - Yeast rho-type GTPase activating protein RGA1/DBM1. - *Caenorhabditis elegans* homeobox protein ceh-14. - *Caenorhabditis elegans* homeobox protein unc-97. - Yeast hypothetical protein YKR090w. - *Caenorhabditis elegans* hypothetical proteins C28H8.6. These proteins generally have two tandem copies of a domain, called LIM (for Lin-11 Isl-1 Mec-3) in their N-terminal section. Zyxin and paxillin are exceptions in that they contain respectively three and four LIM domains at their C-terminal extremity. In apterous, isl-1, LH-2, lin-11, lim-1 to lim-3, lmx-1 and ceh-14 and mec-3 there is a homeobox domain some 50 to 95 amino acids after the LIM domains. In the LIM domain, there are seven conserved cysteine residues and a histidine. The arrangement

322

followed by these conserved residues is C-x(2)-C-x(16,23)-H-x(2)-[CH]-x(2)-C-x(2)-C-x(16,21)-C-x(2,3)-[CHD]. The LIM domain binds two zinc ions [5]. LIM does not bind DNA, rather it seems to act as an interface for protein-protein interaction. A pattern was developed that spans the first half of the LIM domain.

5

Consensus pattern: C-x(2)-C-x(15,21)-[FYWH SEQ ID NO:272]-H-x(2)-[CH]-x(2)-C-x(2)-C-x(3)-[LIVMF SEQ ID NO:2] [The 5 C's and the H bind zinc]

[1] Frey G., Kim S.K., Horvitz H.R. Nature 344:876-879(1990).

10 [2] Baltz R., Evrard J.-L., Domon C., Steinmetz A. Plant Cell 4:1465-1466(1992).

[3] Sanchez-Garcia I., Rabbitts T.H. Trends Genet. 10:315-320(1994).

[4] Mueller A., Xu G., Wells R., Hollenberg C.P., Piepersberg W. Nucleic Acids Res. 22:3151-3154(1994).

[5] Michelsen J.W., Schmeichel K.L., Beckerle M.C., Winge D.R. Proc. Natl. Acad. Sci.

15 U.S.A. 90:4404-4408(1993).

324. (LRR) Leucine Rich Repeat

CAUTION: This Pfam may not find all Leucine Rich Repeats in a protein. Leucine Rich

20 Repeats are short sequence motifs present in a number of proteins with diverse functions and cellular locations. These repeats are usually involved in protein-protein interactions. Each Leucine Rich Repeat is composed of a beta-alpha unit. These units form elongated non-globular structures. Leucine Rich Repeats are often flanked by cysteine rich domains.

Number of members: 3017

25 [1] The leucine-rich repeat: a versatile binding motif. Kobe B, Deisenhofer J; Trends Biochem Sci 1994;19:415-421. [2] Crystal structure of porcine ribonuclease inhibitor, a protein with leucine-rich repeats. Kobe B, Deisenhofer J; Nature 1993;366:751-756.

30 325. Plant lipid transfer protein family signature (LTP)

Plant cells contain proteins, called lipid transfer proteins (LTP) [1,2,3], which are able to facilitate the transfer of phospholipids and other lipids across membranes. These proteins, whose subcellular location is not yet known, could play a major role in membrane biogenesis

by conveying phospholipids such as waxes or cutin from their site of biosynthesis to membranes unable to form these lipids. Plant LTP's are proteins of about 9 Kd (90 amino acids) which contain eight conserved cysteine residues all involved in disulfide bridges, as shown in the following schematic representation.

```

5      +-----+ | +-----+ |||| *****
xCxxxxCxxxxxxCCxxxxxxxxxCxCxxxxxxxxxxCxxxxxCxx ||| +-----|-----+ | +---
-----+

```

'C': conserved cysteine involved in a disulfide bond.

'*': position of the pattern.

10 Consensus pattern: [LIVM SEQ ID NO:4)]-[PA]-x(2)-C-x-[LIVM SEQ ID NO:4)]-x-[LIVM SEQ ID NO:4)]-x-[LIVMFY SEQ ID NO:18)]-x-[LIVM SEQ ID NO:4)]- [ST]-x(3)-[DN]-C-x(2)-[LIVM SEQ ID NO:4)] [The two C's are involved in disulfide bonds]

15 [1] Wirtz K.W.A. Annu. Rev. Biochem. 60:73-99(1991).

[2] Arondel V., Kader J.C. Experientia 46:579-585(1990).

[3] Ohlrogge J.B., Browse J., Somerville C.R. Biochim. Biophys. Acta 1082:1-26(1991).

20 326. (LAMP) Lysosome-associated membrane glycoproteins signatures

Lysosome-associated membrane glycoproteins (lamp) [1] are integral membrane proteins, specific to lysosomes, and whose exact biological function is not yet clear. Structurally, the lamp proteins consist of two internally homologous lysosome-luminal domains separated by a proline-rich hinge region; at the C-terminal extremity there is a transmembrane region followed by a very short cytoplasmic tail. In each of the duplicated domains, there are two conserved disulfide bonds. This structure is schematically represented in the figure below. +--

```

---+ +-----+ +-----+ +-----+ |||||
xCxxxxxCxxxxxxxxxxxxxCxxxxxCxxxxxxxxxCxxxxxCxxxxxxxxxCxxxxxCxxxxxxx <-
-----><Hinge><-----><TM><C>

```

30 closely related types of lamp: lamp-1 and lamp-2. In chicken lamp-1 is known as LEP100. The macrophage protein CD68 (or macrosialin) [2] is a heavily glycosylated integral membrane protein whose structure consists of a mucin-like domain followed by a proline-rich hinge; a single lamp-like domain; a transmembrane region and a short cytoplasmic tail. Two

signature patterns for this family of proteins were developed. The first one is centered on the first conserved cysteine of the duplicated domains. The second corresponds to a region that includes the extremity of the second domain, the totality of the transmembrane region and the cytoplasmic tail.

5

Consensus pattern: [STA]-C-[LIVM SEQ ID NO:4)]-[LIVMFYW SEQ ID NO:26)]-A-x-[LIVMFYW SEQ ID NO:26)]-x(3)-[LIVMFYW SEQ ID NO:26)]- x(3)-Y [C is involved in a disulfide bond] –

10

Consensus pattern: C-x(2)-D-x(3,4)-[LIVM SEQ ID NO:4)](2)-P-[LIVM SEQ ID NO:4)]-x-[LIVM SEQ ID NO:4)]-G-x(2)-[LIVM SEQ ID NO:4)]- x-G-[LIVM SEQ ID NO:4)](2)-x-[LIVM SEQ ID NO:4)](4)-A-[FY]-x-[LIVM SEQ ID NO:4)]-x(2)-[KR]-[RH]- x(1,2)-[STAG SEQ ID NO:20)](2)-Y-[EQ] [C is involved in a disulfide bond]

[1] Fukuda M. J. Biol. Chem. 266:21327-21330(1991).

15

[2] Holness C.L., da Silva R.P., Fawcett J., Gordon S., Simmons D.L. J. Biol. Chem. 268:9661-9666(1993).

327. Lipolytic enzymes "G-D-S-L" family, serine active site

20

Recently [1], a family of lipolytic enzymes has been characterized. This family currently consist of the following proteins:

- *Aeromonas hydrophila* lipase/phosphatidylcholine-sterol acyltransferase.

- *Xenorhabdus luminescens* lipase 1.

- *Vibrio mimicus* arylesterase.

25

- *Escherichia coli* acyl-coA thioesterase I (gene tesA).

- *Vibrio parahaemolyticus* thermolabile hemolysin/atypical phospholipase.

- Rabbit phospholipase AdRab-B, an intestinal brush border protein with esterase and phospholipase A/lysophospholipase activity that could be involved in the uptake of dietary lipids. AdRab-B contains four repeats of about 320 amino acids.

30

- *Arabidopsis thaliana* and *Brassic napus* anther-specific proline-rich protein APG.

- A *Pseudomonas putida* hypothetical protein in trpE-trpG intergenic region. A serine has been identified a part of the active site in the *Aeromonas*, *Vibrio mimicus* and *Escherichia*

coli enzymes. It is located in a conserved sequence motif that can be used as a signature pattern for these proteins.

-Consensus pattern: [LIVMFYAG SEQ ID NO:351]](4)-G-D-S-[LIVM SEQ ID NO:4]]-
 5 x(1,2)-[TAG]-G
 [S is the active site residue]

328. (Lipoprotein 4) Prokaryotic membrane lipoprotein lipid attachment site

- 10 In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signalpeptidase II). The peptidase recognizes a conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]): - Major outer membrane lipoprotein
- 15 (murein-lipoproteins) (gene lpp). - Escherichia coli lipoprotein-28 (gene nlpA). - Escherichia coli lipoprotein-34 (gene nlpB). - Escherichia coli lipoprotein nlpC. - Escherichia coli lipoprotein nlpD. - Escherichia coli osmotically inducible lipoprotein B (gene osmB). - Escherichia coli osmotically inducible lipoprotein E (gene osmE). - Escherichia coli peptidoglycan-associated lipoprotein (gene pal). - Escherichia coli rare lipoproteins A and B
- 20 (genes rplA and rplB). - Escherichia coli copper homeostasis protein cutF (or nlpE). - Escherichia coli plasmids traT proteins. - Escherichia coli Col plasmids lysis proteins. - A number of Bacillus beta-lactamases. - Bacillus subtilis periplasmic oligopeptide-binding protein (gene oppA). - Borrelia burgdorferi outer surface proteins A and B (genes ospA and ospB). - Borrelia hermsii variable major protein 21 (gene vmp21) and 7 (gene vmp7). -
- 25 Chlamydia trachomatis outer membrane protein 3 (gene omp3). - Fibrobacter succinogenes endoglucanase cel-3. - Haemophilus influenzae proteins Pal and Pcp. - Klebsiella pullulunase (gene pulA). - Klebsiella pullulunase secretion protein pulS. - Mycoplasma hyorhinis protein p37. - Mycoplasma hyorhinis variant surface antigens A, B, and C (genes vlpABC). -
- 30 Neisseria outer membrane protein H.8. - Pseudomonas aeruginosa lipopeptide (gene lppL). - Pseudomonas solanacearum endoglucanase egl. - Rhodospseudomonas viridis reaction center cytochrome subunit (gene cytC). - Rickettsia 17 Kd antigen. - Shigella flexneri invasion plasmid proteins mxiJ and mxiM. - Streptococcus pneumoniae oligopeptide transport protein A (gene amiA). - Treponema pallidum 34 Kd antigen. - Treponema pallidum membrane

protein A (gene tmpA). - *Vibrio harveyi* chitobiase (gene chb). - *Yersinia* virulence plasmid protein yscJ. - Halocyanin from *Natrobacterium pharaonis* [4], a membrane associated copper-binding protein. This is the first archaeobacterial protein known to be modified in such a fashion). From the precursor sequences of all these proteins, a consensus pattern and a set of rules to identify this type of post-translational modification was derived.

Consensus pattern: {DERK SEQ ID NO:354}(6)-[LIVMFWSTAG SEQ ID NO:352](2)-[LIVMFYSTAGCQ SEQ ID NO:353]-[AGS]-C [C is the lipid attachment site] Additional rules: 1) The cysteine must be between positions 15 and 35 of the sequence in consideration.
2) There must be at least one Lys or one Arg in the first seven positions of the sequence.

[1] Hayashi S., Wu H.C. J. Bioenerg. Biomembr. 22:451-471(1990).

[2] Klein P., Somorjai R.L., Lau P.C.K. Protein Eng. 2:15-20(1988).

[3] von Heijne G. Protein Eng. 2:531-534(1989).

[4] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol. Chem. 269:14939-14945(1994).

329. (Lipoprotein 5) Prokaryotic membrane lipoprotein lipid attachment site. In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signal peptidase II). The peptidase recognizes a conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]): - Major outer membrane lipoprotein (murein-lipoproteins) (gene lpp). - *Escherichia coli* lipoprotein-28 (gene nlpA). - *Escherichia coli* lipoprotein-34 (gene nlpB). - *Escherichia coli* lipoprotein nlpC. - *Escherichia coli* lipoprotein nlpD. - *Escherichia coli* osmotically inducible lipoprotein B (gene osmB). - *Escherichia coli* osmotically inducible lipoprotein E (gene osmE). - *Escherichia coli* peptidoglycan-associated lipoprotein (gene pal). - *Escherichia coli* rare lipoproteins A and B (genes rplA and rplB). - *Escherichia coli* copper homeostasis protein cutF (or nlpE). - *Escherichia coli* plasmids traT proteins. - *Escherichia coli* Col plasmids lysis proteins. - A number of *Bacillus* beta-lactamases. - *Bacillus subtilis* periplasmic oligopeptide-binding protein (gene oppA). - *Borrelia burgdorferi* outer surface proteins A and B (genes ospA and ospB). - *Borrelia*

hermsii variable major protein 21 (gene vmp21) and 7 (gene vmp7). - Chlamydia trachomatis
 outer membrane protein 3 (gene omp3). - Fibrobacter succinogenes endoglucanase cel-3. -
 Haemophilus influenzae proteins Pal and Pcp. - Klebsiella pullulunase (gene pulA). -
 Klebsiella pullulunase secretion protein pulS. - Mycoplasma hyorhinis protein p37. -
 5 Mycoplasma hyorhinis variant surface antigens A, B, and C (genes vlp ABC). - Neisseria
 outer membrane protein H.8. - Pseudomonas aeruginosa lipopeptide (gene lppL). -
 Pseudomonas solanacearum endoglucanase egl. - Rhodopseudomonas viridis reaction center
 cytochrome subunit (gene cytC). - Rickettsia 17 Kd antigen. - Shigella flexneri invasion
 plasmid proteins mxiJ and mxiM. - Streptococcus pneumoniae oligopeptide transport protein
 10 A (gene amiA). - Treponema pallidum 34 Kd antigen. - Treponema pallidum membrane
 protein A (gene tmpA). - Vibrio harveyi chitobiase (gene chb). - Yersinia virulence plasmid
 protein yscJ. - Halocyanin from Natrobacterium pharaonis [4], a membrane associated
 copper-binding protein. This is the first archaeobacterial protein known to be modified in such
 a fashion). From the precursor sequences of all these proteins, a consensus pattern and a set of
 15 rules to identify this type of post-translational modification have been developed.

Consensus pattern: {DERK SEQ ID NO:354}}(6)-[LIVMFWSTAG SEQ ID NO:352]](2)-
 [LIVMFYSTAGCQ SEQ ID NO:353]]-[AGS]-C [C is the lipid attachment site] Additional
 rules: 1) The cysteine must be between positions 15 and 35 of the sequence in consideration.
 20 2) There must be at least one Lys or one Arg in the first seven positions of the sequence.

[1] Hayashi S., Wu H.C. J. Bioenerg. Biomembr. 22:451-471(1990).[2] Klein P., Somorjai
 R.L., Lau P.C.K. Protein Eng. 2:15-20(1988).[3] von Heijne G. Protein Eng. 2:531-
 534(1989).[4] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M.
 25 J. Biol. Chem. 269:14939-14945(1994).

330. (Lum binding) Riboflavin synthase alpha chain family Lum-binding site signature
 The following proteins have been shown [1,2] to be structurally and evolutionary related: -
 30 Riboflavin synthase alpha chain (RS-alpha) (gene ribC in Escherichia coli, ribB in Bacillus
 subtilis and Photobacterium leiognathi, RIB5 in yeast). This enzyme synthesizes riboflavin
 from two moles of 6,7- dimethyl-8-(1'-D-ribityl)lumazine (Lum), a pteridine-derivative. -
 Photobacterium phosphoreum lumazine protein (LumP) (gene luxL). LumP is a protein that

modulates the color of the bioluminescence emission of bacterial luciferase. In the presence of LumP, light emission is shifted to higher energy values (shorter wavelength). LumP binds non-covalently to 6,7-dimethyl-8-(1'-D-ribityl) lumazine. - *Vibrio fischeri* yellow fluorescent protein (YFP) (gene luxY). Like LumP, YFP modulates light emission but towards a longer wavelength. YFP binds non-covalently to FMN. These proteins seem to have evolved from the duplication of a domain of about 100 residues. In its C-terminal section, this domain contains a conserved motif [KR]-V-N-[LI]-E which has been proposed to be the binding site for Lum.RS-alpha which binds two molecules of Lum has two perfect copies of this motif, while LumP which binds one molecule of Lum, has a Glu instead of Lys/Arg in the first position of the second copy of the motif. Similarly, YFP, which binds to one molecule of FMN, also seems to have a potentially dysfunctional binding site by substitution of Gly for Glu in the last position of the first copy of the motif. Our signature pattern includes the Lum-binding motif.

Consensus pattern: [LIVMF SEQ ID NO:2)]-x(5)-G-[STADNQ SEQ ID NO:355)]-[KREQIYW SEQ ID NO:356)]-V-N-[LIVM SEQ ID NO:4)]-E

[1] O'Kane D.J., Woodward B., Lee J., Prasher D.C. Proc. Natl. Acad. Sci. U.S.A. 88:1100-1104(1991).

[2] O'Kane D.J., Prasher D.C. Mol. Microbiol. 6:443-449(1992).

331. Lysyl oxidase putative copper-binding region signature

Lysyl oxidase (LOX) [1] is an extracellular copper-dependent enzyme that catalyzes the oxidative deamination of peptidyl lysine residues in precursors of various collagens and elastins. The deaminated lysines are then able to form aldehyde cross-links. LOX binds a single copper atom which seems to reside within an octahedral coordination complex which includes at least three histidine ligands. Four histidine residues are clustered in a central region of the enzyme. This region is thought to be involved in copper-binding and is called the 'copper-talon' [1]. This region was used as a signature pattern.

Consensus pattern: W-E-W-H-S-C-H-Q-H-Y-H

[1] Krebs C.J., Krawetz S.A. Biochim. Biophys. Acta 1202:7-12(1993).

332. Metallo-beta-lactamase superfamily (lactamase_B)

- 5 [1] : Neuwald AF, Liu JS, Lipman DJ, Lawrence CE, Nucleic Acids Res 1997;25:1665-1677. [2] Carfi A, Pares S, Duee E, Galleni M, Duez C, Frere JM, Dideberg O, EMBO J 1995;14:4914-4921.

10 333. L-lactate dehydrogenase active site (ldh1)

- L-lactate dehydrogenase (EC 1.1.1.27) (LDH) [1] catalyzes the reversible NAD-dependent interconversion of pyruvate to L-lactate. In vertebrate muscles and in lactic acid bacteria it represents the final step in anaerobic glycolysis. This tetrameric enzyme is present in prokaryotic and eukaryotic organisms. Invertebrates there are three isozymes of LDH: the M form (LDH-A), found predominantly in muscle tissues; the H form (LDH-B), found in heart muscle and the X form (LDH-C), found only in the spermatozoa of mammals and birds. In birds and crocodilian eye lenses, LDH-B serves as a structural protein and is known as epsilon-crystallin [2]. L-2-hydroxyisocaproate dehydrogenase (EC 1.1.1.-) (L-hicDH) [3] catalyzes the reversible and stereospecific interconversion between 2-ketocarboxylic acids and L-2-hydroxy-carboxylic acids. L-hicDH is evolutionary related to LDH's. As a signature for LDH's a region was selected that includes a conserved histidine which is essential to the catalytic mechanism.
- 15
20

Consensus pattern: [LIVMA SEQ ID NO:30)]-G-[EQ]-H-G-[DN]-[ST] [H is the active site residue] -

25

[1] Abad-Zapatero C., Griffith J.P., Sussman J.L., Rossmann M.G. J. Mol. Biol. 198:445-467(1987).

[2] Hendriks W., Mulders J.W.M., Bibby M.A., Slingsby C., Bloemendal H., de Jong W.W. Proc. Natl. Acad. Sci. U.S.A. 85:7114-7118(1988).

30

[3] Lerch H.-P., Frank R., Collins J. Gene 83:263-270(1989).

Malate dehydrogenase active site signature (ldh2)

Malate dehydrogenase (EC 1.1.1.37) (MDH) [1,2] catalyzes the interconversion of malate to oxaloacetate utilizing the NAD/NADH cofactor system. The enzyme participates in the citric acid cycle and exists in all aerobic organisms. While prokaryotic organisms contains a single form of MDH, in eukaryotic cells there are two isozymes: one which is located in the mitochondrial matrix and the other in the cytoplasm. Fungi and plants also harbor a glyoxysomal form which functions in the glyoxylate pathway. In plants chloroplast there is an additional NADP-dependent form of MDH (EC 1.1.1.82) which is essential for both the universal C3 photosynthesis (Calvin) cycle and the more specialized C4 cycle. As a signature pattern for this enzyme a region was chosen that includes two residues involved in the catalytic mechanism [3]: an aspartic acid which is involved in a proton relay mechanism, and an arginine which binds the substrate.

Consensus pattern: [LIVM SEQ ID NO:4)]-T-[TRKMN SEQ ID NO:357)]-L-D-x(2)-R-[STA]-x(3)-[LIVMFY SEQ ID NO:18)] [D and R are the active site residues]-

[1] McAlister-Henn L. Trends Biochem. Sci. 13:178-181(1988).

[2] Gietl C. Biochim. Biophys. Acta 1100:217-234(1992).

[3] Birktoft J.J., Rhodes G., Banaszak L.J. Biochemistry 28:6065-6081(1989).

[4] Cendrin F., Chroboczek J., Zaccai G., Eisenberg H., Mevarech M. Biochemistry 32:4308-4313(1993).

334. Legume lectins signatures

Leguminous plants synthesize sugar-binding proteins which are called legume lectins [1,2].

These lectins are generally found in the seeds. The exact function of legume lectins is not known but they may be involved in the attachment of nitrogen-fixing bacteria to legumes and in the protection against pathogens. Legume lectins bind calcium and manganese (or other transition metals). Legume lectins are synthesized as precursor proteins of about 230 to 260 amino acid residues. Some legume lectins are proteolytically processed to produce two chains: beta (which corresponds to the N-terminal) and alpha (C-terminal). The lectin concanavalin A (conA) from jack bean is exceptional in that the two chains are transposed and ligated (by formation of a new peptide bond). The N-terminus of mature conA thus corresponds to that of the alpha chain and the C-terminus to the beta chain. Two signature

331

patterns specific to legume lectins have been developed: the first is located in the C-terminal section of the beta chain and contains a conserved aspartic acid residue important for the binding of calcium and manganese; the second one is located in the N-terminal of the alpha chain.

5

Consensus pattern: [LIV]-[STAG SEQ ID NO:20)]-V-[DEQV SEQ ID NO:358)]-[FLI]-D-[ST] [D binds manganese and calcium]-

Consensus pattern: [LIV]-x-[EDQ]-[FYWKR SEQ ID NO:359)]-V-x-[LIVF SEQ ID NO:127)]-G-[LF]-[ST]-

10

[1] Sharon N., Lis H. FASEB J. 4:3198-320(1990).

[2] Lis H., Sharon N. Annu. Rev. Biochem. 55:33-37(1986).

15

335. CoA-ligases (ligases- CoA)

This family includes the CoA ligases Succinyl-CoA synthetase alpha: and beta chains, malate CoA ligase and ATP-citrate lyase. Some members of the family utilise ATP others use GTP.

[1] Wolodko WT, Fraser ME, James MN, Bridger WA, J Biol Chem 1994;269:10883-

20

10890.

336. linker histone H1 and H5 family

Linker histone H1 is an essential component of chromatin structure. H1 links nucleosomes into higher order structures Histone H1 is replaced by histone H5 in some cell types.

25

[1] Ramakrishnan V, Finch JT, Graziano V, Lee PL, Sweet RM, Nature 1993;362:219-223.

30

337. Lipocalin signature (lip1)

Proteins which transport small hydrophobic molecules such as steroids, bilins, retinoids, and lipids share limited regions of sequence homology and a common tertiary structure

architecture [1 to 5]. This is an eight stranded antiparallel beta-barrel with a repeated + 1 topology enclosing a internal ligand binding site [1,3]. The name 'lipocalin' has been proposed [5] for this protein family. Proteins known to belong to this family are listed below (references are only provided for recently determined sequences). - Alpha-1-microglobulin (protein HC), which seems to bind porphyrin. - Alpha-1-acid glycoprotein (orosomucoid), which can bind a remarkable array of natural and synthetic compounds [6]. - Aphrodisin which, in hamsters, functions as an aphrodisiac pheromone. - Apolipoprotein D, which probably binds heme-related compounds. - Beta-lactoglobulin, a milk protein whose physiological function appears to bind retinol. - Complement component C8 gamma chain, which seems to bind retinol [7]. - Crustacyanin [8], a protein from lobster carapace, which binds astaxanthin, a carotenoid. - Epididymal-retinoic acid binding protein (E-RABP) [9] involved in sperm maturation. - Insectacyanin, a moth bilin-binding protein, and a related butterfly bilin- binding protein (BBP). - Late Lactation protein (LALP), a milk protein from tammar wallaby [10]. - Neutrophil gelatinase-associated lipocalin (NGAL) (p25) (SV-40 induced 24p3 protein) [11]. - Odorant-binding protein (OBP), which binds odorants. - Plasma retinol-binding proteins (PRBP). - Human pregnancy-associated endometrial alpha-2 globulin. - Probasin (PB), a rat prostatic protein. - Prostaglandin D synthase (EC 5.3.99.2) (GSH-independent PGD synthetase), a lipocalin with enzymatic activity [12]. - Purpurin, a retinal protein which binds retinol and heparin. - Quiescence specific protein p20K from chicken (embryo CH21 protein). - Rodent urinary proteins (alpha-2-microglobulin), which may bind pheromones. - VNSP 1 and 2, putative pheromone transport proteins from mouse vomeronasal organ [13]. - Von Ebner's gland protein (VEGP) [14] (also called tear lipocalin), a mammalian protein which may be involved in taste recognition. - A frog olfactory protein, which may transport odorants. - A protein found in the cerebrospinal fluid of the toad Bufo Marinus with a supposed function similar to transthyretin in transport across the blood brain barrier [15]. - Lizard's epididymal secretory protein IV (LESP IV), which could transport small hydrophobic molecules into the epididymal fluid during sperm maturation [16]. - Prokaryotic outer-membrane protein blc [17]. The sequences of most members of the family, the core or kernal lipocalins, are characterized by three short conserved stretches of residues [3,18]. Others, the outlier lipocalin group, share only one or two of these [3,18]. A signature pattern was built around the first, common to all outlier and kernallipocalins, which occurs near the start of the first beta-strand.

Consensus pattern: [DENG SEQ ID NO:360)]-x-[DENQGSTARK SEQ ID NO:361)]-x(0,2)-[DENQARK SEQ ID NO:362)]-[LIVFY SEQ ID NO:257)]-{CP}-G-{C}- W-[FYWLRH SEQ ID NO:363)]-x-[LIVMTA SEQ ID NO:311)]-

Note: it is suggested, on the basis of similarities of structure, function, and sequence, that this family forms an overall superfamily, called the calycins, with the avidin/streptavidin <PDOC00499> and the cytosolic fatty- acid binding proteins <PDOC00188> families [3,19]

[1] Cowan S.W., Newcomer M.E., Jones T.A. Proteins 8:44-61(1990).

[2] Igarashi M., Nagata A., Toh H., Urade H., Hayaishi N. Proc. Natl. Acad. Sci. U.S.A. 89:5376-5380(1992).

[3] Flower D.R., North A.C.T., Attwood T.K. Protein Sci. 2:753-761(1993).

[4] Godovac-Zimmermann J. Trends Biochem. Sci. 13:64-66(1988).

[5] Pervaiz S., Brew K. FASEB J. 1:209-214(1987).

[6] Kremer J.M.H., Wilting J., Janssen L.H.M. Pharmacol. Rev. 40:1-47(1989).

[7] Haefliger J.-A., Peitsch M.C., Jenne D., Tschopp J. Mol. Immunol. 28:123-131(1991).

[8] Keen J.N., Caceres I., Eliopoulos E.E., Zagalsky P.F., Findlay J.B.C. Eur. J. Biochem. 197:407-417(1991).

[9] Newcomer M.E. Structure 1:7-18(1993).

[10] Collet C., Joseph R. Biochim. Biophys. Acta 1167:219-222(1993).

[11] Kjeldsen L., Johnsen A.H., Sengelov H., Borregaard N. J. Biol. Chem. 268:10425-10432(1993).

[12] Peitsch M.C., Boguski M.S. Trends Biochem. Sci. 16:363-363(1991).

[13] Miyawaki A., Matsushita Y.R., Ryo Y., Mikoshiba T. EMBO J. 13:5835-5842(1994).

[14] Kock K., Ahlers C., Schmale H. Eur. J. Biochem. 221:905-916(1994).

[15] Achen M.G., Harms P.J., Thomas T., Richardson S.J., Wettenhall R.E.H., Schreiber G. J. Biol. Chem. 267:23170-23174(1992).

[16] Morel L., Dufarre J.-P., Depeiges A. J. Biol. Chem. 268:10274-10281(1993).

[17] Bishop R.E., Penfold S.S., Frost L.S., Holtje J.V., Weiner J.H. J. Biol. Chem. 270:23097-23103(1995).

[18] Flower D.R., North A.C.T., Attwood T.K. Biochem. Biophys. Res. Commun. 180:69-74(1991).

[19] Flower D.R. FEBS Lett. 333:99-102(1993).

Cytosolic fatty-acid binding proteins signature (lip2)

A number of low molecular weight proteins which bind fatty acids and other organic anions are present in the cytosol [1,2]. Most of them are structurally related and have probably diverged from a common ancestor. This structure is a ten stranded antiparallel beta-barrel, albeit with a wide discontinuity between the fourth and fifth strands, with a repeated +1 topology enclosing an internal ligand binding site [2,7]. Proteins known to belong to this family include: - Six, tissue-specific, types of fatty acid binding proteins (FABPs) found in liver, intestine, heart, epidermal, adipocyte, brain/retina. Heart FABP is also known as mammary-derived growth inhibitor (MDGI), a protein that reversibly inhibits proliferation of mammary carcinoma cells. Epidermal FABP is also known as psoriasis-associated FABP [3]. - Insect muscle fatty acid-binding proteins. - Testis lipid binding protein (TLBP). - Cellular retinol-binding proteins I and II (CRBP). - Cellular retinoic acid-binding protein (CRABP). - Gastrotropin, an ileal protein which stimulates gastric acid and pepsinogen secretion. It seems that gastrotropin binds to bile salts and bilirubins. - Fatty acid binding proteins MFB1 and MFB2 from the midgut of the insect *Manduca sexta* [4]. In addition to the above cytosolic proteins, this family also includes: - Myelin P2 protein, which may be a lipid transport protein in Schwann cells. P2 is associated with the lipid bilayer of myelin. - *Schistosoma mansoni* protein Sm14 [5] which seems to be involved in the transport of fatty acids. - *Ascaris suum* p18 a secreted protein that may play a role in sequestering potentially toxic fatty acids and their peroxidation products or that may be involved in the maintenance of the impermeable lipid layer of the eggshell. - Hypothetical fatty acid-binding proteins F40F4.2, F40F4.3, F40F4.4 and ZK742.5 from *Caenorhabditis elegans*. As a signature pattern for these proteins a segment from the N-terminal extremity was used.

Consensus pattern: [GSAIVK SEQ ID NO:364]-x-[FYW]-x-[LIVMF SEQ ID NO:2])-x(4)-[NHG]-[FY]-[DE]-x- [LIVMFY SEQ ID NO:18)]-[LIVM SEQ ID NO:4)]-x(2)-[LIVMAKR SEQ ID NO:365)]-

Note: it is suggested, on the basis of similarities of structure, function, and sequence, that this family forms an overall superfamily, called the calycins, with the lipocalin <PDOC00187> and avidin/streptavidin <PDOC00499> families [6,7].

[1] Bernier I., Jolles P. Biochimie 69:1127-1152(1987).

[2] Veerkamp J.H., Peeters R.A., Maatman R.G.H.J. Biochim. Biophys. Acta 1081:1-24(1991).

[3] Siegenthaler G., Hotz R., Chatellard-Gruaz D., Didierjean L., Hellman U., Saurat J.-H. Biochem. J. 302:363-371(1994).

5 [4] Smith A.F., Tsuchida K., Hanneman E., Suzuki T.C., Wells M.A. J. Biol. Chem. 267:380-384(1992).

[5] Moser D., Tendler M., Griffiths G., Klinkert M.-Q. J. Biol. Chem. 266:8447-8454(1991).

[6] Flower D.R., North A.C.T, Attwood T.K. Protein Sci. 2:753-761(1993).

[7] Flower D.R. FEBS Lett. 333:99-102(1993).

10

338. Lipoxygenases iron-binding region signatures

Lipoxygenases (EC 1.13.11.-) are a class of iron-containing dioxygenases which catalyzes the hydroperoxidation of lipids, containing a cis,cis-1,4-pentadiene structure. They are common

15 in plants where they may be involved in a number of diverse aspects of plant physiology including growth and development, pest resistance, and senescence or responses to wounding [1]. In mammals a number of lipoxygenases isozymes are involved in the metabolism of prostaglandins and leukotrienes [2]. Sequence data is available for the following

20 lipoxygenases: - Plant lipoxygenases (EC 1.13.11.12). Plants express a variety of cytosolic isozymes as well as what seems [3] to be a chloroplast isozyme. - Mammalian arachidonate 5-lipoxygenase (EC 1.13.11.34). - Mammalian arachidonate 12-lipoxygenase (EC 1.13.11.31). - Mammalian erythroid cell-specific 15-lipoxygenase (EC 1.13.11.33). The iron atom in lipoxygenases is bound by four ligands, three of which are histidine residues [4]. Six histidines are conserved in all lipoxygenase sequences, five of them are found clustered in a

25 stretch of 40 amino acids. This region contains two of the three zinc-ligands; the other histidines have been shown [5] to be important for the activity of lipoxygenases. As signatures for this family of enzymes two patterns in the region of the histidine cluster were selected. The first pattern contains the first three conserved histidines and the second pattern includes the fourth and the fifth.

30

Consensus pattern: H-[EQ]-x(3)-H-x-[LM]-[NQRCS SEQ ID NO:366)]-[GST]-H-[LIVMSTAC SEQ ID NO:151)](3)-E [The second and third H's bind iron]-

Consensus pattern: [LIVMA SEQ ID NO:30)]-H-P-[LIVM SEQ ID NO:4)]-x-[KRQ]-
[LIVMF SEQ ID NO:2)](2)-x-[AP]-H-

- [1] Vick B.A., Zimmerman D.C. (In) Biochemistry of plants: A comprehensive treatise,
5 Stumpf P.K., Ed., Vol. 9, pp.53-90, Academic Press, New-York, (1987).
[2] Needleman P., Turk J., Jakschik B.A., Morrison A.R., Lefkowitz J.B. Annu. Rev.
Biochem. 55:69-102(1986).
[3] Peng Y.L., Shirano Y., Ohta H., Hibino T., Tanaka K., Shibata D. J. Biol. Chem.
269:3755-3761(1994).
10 [4] Boyington J.C., Gaffney B.J., Amzel L.M. Science 260:1482-1486(1993).
[5] Steczko J., Donoho G.P., Clemens J.C., Dixon J.E., Axelrod B. Biochemistry 31:4053-
4057(1992).

15 339. Fumarate lyases signature (lyase_1)

A number of enzymes, belonging to the lyase class, for which fumarate is a substrate have
been shown [1,2] to share a short conserved sequence around a methionine which is probably
involved in the catalytic activity of this type of enzymes. These enzymes are: - Fumarase (EC
4.2.1.2) (fumarate hydratase), which catalyzes the reversible hydration of fumarate to L-
20 malate. There seem to be 2 classes of fumarases: class I are thermolabile dimeric enzymes (as
for example: Escherichia coli fumC); class II enzymes are thermostable and tetrameric and
are found in prokaryotes (as for example: Escherichia coli fumA and fumB) as well as in
eukaryotes. The sequence of the two classes of fumarases are not closely related. - Aspartate
ammonia-lyase (EC 4.3.1.1) (aspartase), which catalyzes the reversible conversion of
25 aspartate to fumarate and ammonia. This reaction is analogous to that catalyzed by fumarase,
except that ammonia rather than water is involved in the trans-elimination reaction. -
Arginosuccinase (EC 4.3.2.1) (argininosuccinate lyase), which catalyzes the formation of
arginine and fumarate from arginosuccinate, the last step in the biosynthesis of arginine. -
Adenylosuccinase (EC 4.3.2.2) (adenylosuccinate lyase) [3], which catalyzes the eight step in
30 the de novo biosynthesis of purines, the formation of 5'-phosphoribosyl-5-amino-4-
imidazolecarboxamide and fumarate from 1-(5-phosphoribosyl)-4-(N-succino-carboxamide).
That enzyme can also catalyzes the formation of fumarate and AMP from adenylosuccinate. -
Pseudomonas putida 3-carboxy-cis,cis-muconate cycloisomerase (EC 5.5.1.2) (3-

carboxymuconate lactonizing enzyme) (gene *pcaB*) [4], an enzyme involved in aromatic acids catabolism

Consensus pattern: G-S-x(2)-M-x(2)-K-x-N-

5

[1] Woods S.A., Shwartzbach S.D., Guest J.R. *Biochim. Biophys. Acta* 954:14-26(1988).

[2] Woods S.A., Miles J.S., Guest J.R. *FEMS Microbiol. Lett.* 51:181-186(1988).

[3] Zalkin H., Dixon J.E. *Prog. Nucleic Acid Res. Mol. Biol.* 42:259-287(1992).

[4] Williams S.E., Woolridge E.M., Ransom S.C., Landro J.A., Babbitt P.C., Kozarich J.W.

10 *Biochemistry* 31:9768-9776(1992).

340. MCM family signature and profile

Proteins shown to be required for the initiation of eukaryotic DNA replication share a highly
15 conserved domain of about 210 amino-acid residues [1,2,3]. The latter shows some
similarities [4] with that of various other families of DNA-dependent ATPases. Eukaryotes
seem to possess a family of six proteins that contain this domain. They were first identified in
yeast where most of them have a direct role in the initiation of chromosomal DNA replication
by interacting directly with autonomously replicating sequences (ARS). They were thus
20 called 'minichromosome maintenance proteins' with gene symbols prefixed by MCM. These
six proteins are: - MCM2, also known as *cdc19* (in *S.pombe*) [E1]. - MCM3, also known as
DNA polymerase alpha holoenzyme-associated protein P1, RLF beta subunit or ROA. -
MCM4, also known as *CDC54*, *cdc21* (in *S.pombe*) or *dpa* (in *Drosophila*). - MCM5, also
known as *CDC46* or *nda4* (in *S.pombe*). - MCM6, also known as *mis5* (in *S.pombe*). -
25 MCM7, also known as *CDC47* or *Prolifera* (in *A.thaliana*). This family is also present in
archaea. In *Methanococcus jannaschii* there are four members: MJ0363, MJ0961,
MJ1489 and MJECL13. The presence of a putative ATP-binding domain implies that these
proteins maybe involved in an ATP-consuming step in the initiation of DNA replication in
eukaryotes. As a signature pattern, a perfectly conserved region was selected that represents a
30 special version of the B motif found in ATP-binding proteins.

Consensus pattern: G-[IVT]-[LVAC SEQ ID NO:367]](2)-[IVT]-D-[DE]-[FL]-[DNST SEQ
ID NO:265]]

- [1] Coxon A., Maundrell K., Kearsey S.E. Nucleic Acids Res. 20:5571-5577(1992).
[2] Hu B., Burkhardt R., Schulte D., Musahl C., Knippers R. Nucleic Acids Res. 21:5289-5293(1993).
5 [3] Tye B.-K. Trends Cell Biol. 4:160-166(1994).
[4] Koonin E.V. Nucleic Acids Res. 21:2541-2547(1993).

341. Macrophage migration inhibitory factor family signature (MIF)

10 A protein called macrophage migration inhibitory factor (MIF) [1] seems to exert an important role in host inflammatory responses. It play a pivotal role in the host response to endotoxic shock and appears to serve as a pituitary "stress" hormone that regulates systemic inflammatory responses. MIF is a secreted protein of 115 residues which is not processed from a larger precursor. D-dopachrome tautomerase [2] is a mammalian cytoplasmic enzyme
15 involved in melanin biosynthesis and that tautomerizes D-dopachrome with concomitant decarboxylation to give 5,6-dihydroxyindole (DHI). It is a protein of 117 residues highly related to MIF. It must be noted that MIF binds glutathione and has been said to be related to glutathione S-transferases. This assertion has been later disproved [3].As a signature pattern for these proteins, a conserved region was selected located in the central section.

20

Consensus pattern: [DE]-P-C-A-x(3)-[LIVM SEQ ID NO:4)]-x-S-I-G-x-[LIVM SEQ ID NO:4)]-G-

25

- [1] Bucala R. Immunol. Lett. 43:23-26(1994).
[2] Odh G., Hindemith A., Rosengren A.-M., Rosengren E., Rorsman H. Biochem. Biophys. Res. Commun. 197:619-624(1993).
[3] Pearson W.R. Protein Sci. 3:525-527(1994).

30

342. MIP family signature

Recently the sequence of a number of different proteins, that all seem to be transmembrane channel proteins, has been found to be highly related [1 to 4].These proteins are listed below.

- Mammalian major intrinsic protein (MIP). MIP is the major component of lens fiber gap

junctions. Gap junctions mediate direct exchange of ions and small molecule from one cell to another. - Mammalian aquaporins [5]. These proteins form water-specific channels that provide the plasma membranes of red cells and kidney proximal and collecting tubules with high permeability to water, thereby permitting water to move in the direction of an osmotic gradient. - Soybean nodulin-26, a major component of the peribacteroid membrane induced during nodulation in legume roots after *Rhizobium* infection. - Plants tonoplast intrinsic proteins (TIP). There are various isoforms of TIP: alpha (seed), gamma, Rt (root), and Wsi (water-stress induced). These proteins may allow the diffusion of water, amino acids and/or peptides from the tonoplast interior to the cytoplasm. - Bacterial glycerol facilitator protein (gene *glpF*), which facilitates the movement of glycerol across the cytoplasmic membrane. - *Salmonella typhimurium* propanediol diffusion facilitator (gene *pduF*). - Yeast FPS1, a glycerol uptake/efflux facilitator protein. - *Drosophila* neurogenic protein 'big brain' (*bib*). This protein may mediate intercellular communication; it may functions by allowing the transport of certain molecule(s) and thereby sending a signal for an exodermal cell to become an epidermoblast instead of a neuroblast. - Yeast hypothetical protein YFL054c. - A hypothetical protein from the *pepX* region of *Lactococcus lactis*. The MIP family proteins seem to contain six transmembrane segments. Computer analysis shows that these protein probably arose by a tandem, intragenic duplication event from an ancestral protein that contained three transmembrane segments. As a signature pattern a well conserved region was selected which is located in a probable cytoplasmic loop between the second and third transmembrane regions.

Consensus pattern: [HNQA SEQ ID NO:368)]-x-N-P-[STA]-[LIVMF SEQ ID NO:2)]-[ST]-[LIVMF SEQ ID NO:2)]-[GSTAFY SEQ ID NO:369)]-

[1] Reizer J., Reizer A., Saier M.H. Jr. *CRC Crit. Rev. Biochem.* 28:235-257(1993).

[2] Baker M.E., Saier M.H. Jr. *Cell* 60:185-186(1990).

[3] Pao G.M., Wu L.-F., Johnson K.D., Hoefte H., Chrispeels M.J., Sweet G., Sandal N.N., Saier M.H. Jr. *Mol. Microbiol.* 5:33-37(1991).

[4] Wistow G.J., Pisano M.M., Chepelinsky A.B. *Trends Biochem. Sci.* 16:170-171(1991).

[5] Chrispeels M.J., Agre P. *Trends Biochem. Sci.* 19:421-425(1994).

343. Mandelate racemase / muconate lactonizing enzyme family signatures

Mandelate racemase (EC 5.1.2.2) (MR) and muconate lactonizing enzyme (EC 5.5.1.1) (MLE) are two bacterial enzymes involved in aromatic acid catabolism. They catalyze mechanistically distinct reactions yet they are related at the level of their primary, quaternary (homooctamer) and tertiary structures [1,2]. A number of other proteins also seem to be evolutionary related to these two enzymes. These are: - The various plasmid-encoded chloromuconate cycloisomerases (EC 5.5.1.7). - *Escherichia coli* protein *rspA* [3], *rspA* seems to be involved in the degradation of homoserine lactone (HSL) or of one of its metabolite. - *Escherichia coli* hypothetical protein *ycjG*. - *Escherichia coli* hypothetical protein *ycjU*. - A hypothetical protein from *Streptomyces ambofaciens* [4]. Two signature patterns have been developed for these enzymes; both contain conserved acidic residues. The second pattern contains an aspartate and a glutamate which are ligands for either a magnesium ion (in MR) or a manganese ion (in MLE).

Consensus pattern: A-x-[SAGCN SEQ ID NO:370)]-[SAG]-[LIVM SEQ ID NO:4)]-[DEQ]-x-A-[LA]-x-[DE]-[LIA]-x-[GA]-[KRQ]-x(4)-[PSA]-[LIV]-x(2)-L-[LIVMF SEQ ID NO:2)]-G-

Consensus pattern: [LIVF SEQ ID NO:127)]-x(2)-D-x-[NH]-x(7)-[ACL]-x(6)-[LIVMF SEQ ID NO:2)]-x(7)-[LIVM SEQ ID NO:4)]-E-[DENQ SEQ ID NO:371)]-P [D and E bind a divalent metal ion]-

[1] Neidhart D.J., Kenyon G.L., Gerlt J.A., Petsko G.A. *Nature* 347:692-694(1990).

[2] Petsko G.A., Kenyon G.L., Gerlt J.A., Ringe D., Kozarich J.W. *Trends Biochem. Sci.* 18:372-376(1993).

[3] Huisman G.W., Kolter R. *Science* 265:537-539(1994).

[4] Schneider D., Aigle B., Leblond P., Simonet J.M., Decaris B. J. *Gen. Microbiol.* 139:2559-2567(1993).

344. Merozoite Surface Antigen 2 (MSA-2) family

Thomas AW, Carr DA, Carter JM, Lyon JA, *Mol Biochem Parasitol* 1990;43:211-220.

345. MSP (Major sperm protein) domain.

Major sperm proteins are involved in sperm motility. These proteins oligomerise to form filaments. Partial matches to this domain are also found in other non MSP proteins.

5 These include [Swiss:P40075](#) and [Swiss:P34593](#).

[1] Bullock TL, Roberts TM, Stewart M, J Mol Biol 1996;263:284-296. [2] King KL, Stewart M, Roberts TM, Seavy M, J Cell Sci 1992;101:847-857.

10 346. (Matrix) Viral matrix protein. Found in Morbillivirus and paramyxovirus, pneumovirus.
Number of members: 105

347. O-methyltransferase (methyltransf)

15 This family includes a range of O-methyltransferases. These enzymes utilise S-adenosyl methionine.

[1] Keller NP, Dischinger HC, Bhatnagar D, Cleveland TE, Ullah AH, Appl Environ Microbiol 1993;59:479-484.

20

348. Magnesium chelatase, subunit ChII

Magnesium-chelatase is a three-component enzyme that catalyses the insertion of Mg²⁺ into protoporphyrin IX. This is the first unique step in the synthesis of (bacterio)chlorophyll. Due to this, it is thought that Mg-chelatase has an important role in channeling inter- mediates into the (bacterio)chlorophyll branch in response to conditions
25 suitable for photosynthetic growth. ChII and BchD have molecular weight between 38-42 kDa.

[1] Walker CJ, Willows RD, Biochem J 1997;327:321-333. [2] Petersen BL, Jensen PE, Gibson LC, Stummann BM, Hunter CN, Henningsen KW, J Bacteriol 1998;180:699-704.

30

349. Plasmid recombination enzyme (Mob_Pre)

With some plasmids, recombination can occur in a site specific manner that is independent of RecA. In such cases, the recombination event requires another protein called Pre. Pre is a plasmid recombination enzyme. This protein is: also known as Mob (conjugative mobilization).

[1] Priebe SD, Lacks SA, J Bacteriol 1989;171:4778-4784.

350. Monooxygenase

This family includes diverse enzymes that utilise FAD.

[1] Gatti DL, Palfey BA, Lah MS, Entsch B, Massey V, Ballou DP, Ludwig ML, Science 1994;266:110-114.

351. Mov34 family

Members of this family are found in proteasome regulatory subunits, eukaryotic initiation factor 3 (eIF3) subunits and regulators of transcription factors.

[1] Aravind L, Ponting CP, Protein Sci 1998;7:1250-1254. [2] Hershey JW, Asano K, Naranda T, Vornlocher HP, Hanachi P, Merrick WC, Biochimie 1996;78:903-907.

352. Myc amino-terminal region (Myc_N_term)

The myc family belongs to the basic helix-loop-helix leucine zipper class of transcription factors, see HLH. Myc forms a heterodimer with Max, and this complex regulates cell growth through direct activation of genes involved in cell replication [2].

[1] Facchini LM, Penn LZ, FASEB J 1998;12:633-651. [2] Grandori C, Eisenman RN, Trends Biochem Sci 1997;22:177-181.

353. (Metallothio_2) Metallothionein. Members of this family are metallothioneins. These proteins are cysteine rich proteins that bind to heavy metals. Members of this family appear to be closest to Class II metallothioneins, seed metalthio. Number of members: 55

[1] Medline: 98267202. Characterization of gene repertoires at mature stage of citrus fruits through random sequencing and analysis of redundant metallothionein-like genes expressed during fruit development. Moriguchi T, Kita M, Hisada S, Endo-Inagaki T, Omura M; Gene 1998;211:221-227.

5

354. MAGE family

The MAGE (melanoma antigen-encoding gene) family are expressed in a wide variety of tumors but not in normal cells, with the exception of the male germ cells, placenta, and, possibly, cells of the developing embryo. The cellular function of this family is unknown.

[1] McCurdy DK, Tai LQ, Nguyen J, Wang Z, Yang HM, Udar N, Naiem F, Concannon P, Gatti RA; Mol Genet Metab 1998;63:3-13.

355. Malic enzymes signature. Malic enzymes, or malate oxidoreductases, catalyze the oxidative decarboxylation of malate into pyruvate important for a wide range of metabolic pathways. There are three related forms of malic enzyme [1,2,3]: - NAD-dependent malic enzyme (EC 1.1.1.38), which uses preferentially NAD and has the ability to decarboxylate oxaloacetate (OAA). It is found in bacteria and insects. - NAD-dependent malic enzyme (EC 1.1.1.39), which uses preferentially NAD and is unable to decarboxylate OAA. It is found in the mitochondrial matrix of plants and is a heterodimer of highly related subunits. - NADP-dependent malic enzyme (EC 1.1.1.40), which has a preference for NADP and has the ability to decarboxylate OAA. This form has been found in fungi, animals and plants. In mammals, there are two isozymes: one, mitochondrial and the other, cytosolic. Plants also have two isozymes: chloroplastic and cytosolic. There are two other proteins which are closely structurally related to malicenzymes: - Escherichia coli protein sfcA, whose function is not yet known but which could be an NAD or NADP-dependent malic enzyme. - Yeast hypothetical protein YKL029c, a probable malic enzyme. There are three well conserved regions in the enzyme sequences. Two of them seem to be involved in binding NAD or

NADP. The significance of the third one, located in the central part of the enzymes, is not yet known. This region has been developed as a signature pattern for these enzymes.

Consensus pattern: F-x-[DV]-D-x(2)-G-T-[GSA]-x-[IV]-x-[LIVMA SEQ ID NO:30)]-
5 [GAST SEQ ID NO:179)](2)- [LIVMF SEQ ID NO:2)](2)-

[1] Artus N.N., Edwards G.E. FEBS Lett. 182:225-233(1985).[2] Loeber G., Infante A.A., Maurer-Fogy I., Krystek E., Dworkin M.B. J. Biol. Chem. 266:3016-3021(1991). [3] Long J.J., Wang J.-L., Berry J.O. J. Biol. Chem. 269:2827-2833(1994).

10 356. (matrixin)

Matrixins cysteine switch (aka peptidase_M10)

15 Mammalian extracellular matrix metalloproteinases (EC 3.4.24.-), also known as matrixins [1] (see <PDOC00129>), are zinc-dependent enzymes. They are secreted by cells in an inactive form (zymogen) that differs from the mature enzyme by the presence of an N-terminal propeptide. A highly conserved octapeptide is found two residues downstream of the
20 C-terminal end of the propeptide. This region has been shown to be involved in autoinhibition of matrixins [2,3]; a cysteine within the octapeptide chelates the active site zinc ion, thus inhibiting the enzyme. This region has been called the 'cysteine switch' or 'autoinhibitor region'.

A cysteine switch has been found in the following zinc proteases:

- 25
- MMP-1 (EC 3.4.24.7) (interstitial collagenase).
 - MMP-2 (EC 3.4.24.24) (72 Kd gelatinase).
 - MMP-3 (EC 3.4.24.17) (stromelysin-1).
 - MMP-7 (EC 3.4.24.23) (matrilysin).
 - 30 - MMP-8 (EC 3.4.24.34) (neutrophil collagenase).
 - MMP-9 (EC 3.4.24.35) (92 Kd gelatinase).
 - MMP-10 (EC 3.4.24.22) (stromelysin-2).
 - MMP-11 (EC 3.4.24.-) (stromelysin-3).

- MMP-12 (EC 3.4.24.65) (macrophage metalloelastase).
- MMP-13 (EC 3.4.24.-) (collagenase 3).
- MMP-14 (EC 3.4.24.-) (membrane-type matrix metalloproteinase 1).
- MMP-15 (EC 3.4.24.-) (membrane-type matrix metalloproteinase 2).
- 5 - MMP-16 (EC 3.4.24.-) (membrane-type matrix metalloproteinase 3).
- Sea urchin hatching enzyme (EC 3.4.24.12) (envelysin) [4].
- Chlamydomonas reinhardtii gamete lytic enzyme (GLE) [5].

10 Consensus pattern P-R-C-[GN]-x-P-[DR]-[LIVSAPKQ SEQ ID NO:372] [C chelates the zinc ion] Sequences known to belong to this class detected by the pattern ALL, except for cat MMP-7 and mouse MMP-11.

[1] Woessner J. Jr. FASEB J. 5:2145-2154(1991).

15 [2] Sanchez-Lopez R., Nicholson R., Gesnel M.C., Matrisian L.M., Breathnach R. J. Biol. Chem. 263:11892-11899(1988).

[3] Park A.J., Matrisian L.M., Kells A.F., Pearson R., Yuan Z., Navre M. J. Biol. Chem. 266:1584-1590(1991).

[4] Lepage T., Gache C. EMBO J. 9:3003-3012(1990).

20 [5] Kinoshita T., Fukuzawa H., Shimada T., Saito T., Matsuda Y. Proc. Natl. Acad. Sci. U.S.A. 89:4693-4697(1992).

357. Vertebrate metallothioneins signature (metalthio)

25 Metallothioneins (MT) [1,2,3] are small proteins which bind heavy metals such as zinc, copper, cadmium, nickel, etc., through clusters of thiolate bonds. MT's occur throughout the animal kingdom and are also found in higher plants, fungi and some prokaryotes. On the basis of structural relationships MT's have been subdivided into three classes. Class I includes mammalian MT's as well as MT's from crustacean and molluscs, but with clearly related

30 primary structure. Class II groups together MT's from various species such as sea urchins, fungi, insects and cyanobacteria which display none or only very distant correspondence to class I MT's. Class III MT's are atypical polypeptides containing gamma-glutamylcysteinyl units. Vertebrate class I MT's are proteins of 60 to 68 amino acid residues, 20 of these

residues are cysteines that bind to 7 bivalent metal ions. As a signature pattern a region that spans 19 residues and which contains seven of the metal-binding cysteines was chosen, this region is located in the N-terminal section of class-I MT's.

5 Consensus pattern: C-x-C-[GSTAP SEQ ID NO:373)]-x(2)-C-x-C-x(2)-C-x-C-x(2)-C-x-K-

[1] Hamer D.H. Annu. Rev. Biochem. 55:913-951(1986).

[2] Kagi J.H.R., Schaffer A. Biochemistry 27:8509-8515(1988).

[3] Binz P.-A. Thesis, 1996, University of Zurich.

10

358. Mitochondrial energy transfer proteins signature (mito_carr)

Different types of substrate carrier proteins involved in energy transfer are found in the inner mitochondrial membrane [1 to 5]. These are: - The ADP,ATP carrier protein (AAC)

15 (ADP/ATP translocase) which exports ATP into the cytosol and imports ADP into the mitochondrial matrix. The sequence of AAC has been obtained from various mammalian, plant and fungal species. - The 2-oxoglutarate/malate carrier protein (OGCP), which exports 2-oxoglutarate into the cytosol and imports malate or other dicarboxylic acids into the mitochondrial matrix. This protein plays an important role in several metabolic processes

20 such as the malate/aspartate and the oxoglutarate/isocitrate shuttles. - The phosphate carrier protein, which transports phosphate groups from the cytosol into the mitochondrial matrix. - The brown fat uncoupling protein (UCP) which dissipates oxidative energy into heat by transporting protons from the cytosol into the mitochondrial matrix. - The tricarboxylate transport protein (or citrate transport protein) which is involved in citrate-H⁺/malate

25 exchange. It is important for the bioenergetics of hepatic cells as it provides a carbon source for fatty acid and sterol biosyntheses, and NAD for the glycolytic pathway. - The Grave's disease carrier protein (GDC), a protein of unknown function recognized by IgG in patients with active Grave's disease. - Yeast mitochondrial proteins MRS3 and MRS4. The exact function of these proteins is not known. They suppress a mitochondrial splice defect in the

30 first intron of the COB gene and may act as carriers, exerting their suppressor activity by modulating solute concentrations in the mitochondrion. - Yeast mitochondrial FAD carrier protein (gene FLX1). - Yeast protein ACR1 [6], which seems essential for acetyl-CoA synthetase activity. - Yeast protein PET8. - Yeast protein PMT. - Yeast protein RIM2. - Yeast

protein YHM1/SHM1. - Yeast protein YMC1. - Yeast protein YMC2. - Yeast hypothetical proteins YBR291c, YEL006w, YER053c, YFR045w, YHR002w, and YIL006w. -

Caenorhabditis elegans hypothetical protein K11H3.3. Two other proteins have been found to belong to this family, yet are not localized in the mitochondrial inner membrane: - Maize amyloplast Brittle-1 protein. This protein, found in the endosperm of kernels, could play a role in amyloplast membrane transport. - Candida boidinii peroxisomal membrane protein PMP47 [7]. PMP47 is an integral membrane protein of the peroxisome and it may play a role as a transporter. These proteins all seem to be evolutionary related. Structurally, they consist of three tandem repeats of a domain of approximately one hundred residues. Each of these domains contains two transmembrane regions. As a signature pattern, one of the most conserved regions in the repeated domain was selected, located just after the first transmembrane region.

Consensus pattern: P-x-[DE]-x-[LIVAT SEQ ID NO:374)]-[RK]-x-[LRH]-[LIVMFY SEQ ID NO:18)]-[QGAIVM SEQ ID NO:375)]-

[1] Klingenberg M. Trends Biochem. Sci. 15:108-112(1990).

[2] Walker J.E. Curr. Opin. Struct. Biol. 2:519-526(1992).

[3] Kuan J., Saier M.H. Jr. CRC Crit. Rev. Biochem. 28:209-233(1993).

[4] Kuan J., Saier M.H. Jr. Res. Microbiol. 144:671-672(1993).

[5] Nelson D.R., Lawson J.E., Klingenberg M., Douglas M.G. J. Mol. Biol. 230:1159-1170(1993).

[6] Palmieri F. FEBS Lett. 346:48-54(1994).

[7] Jank B., Habermann B., Schweyen R.J., Link T.A. Trends Biochem. Sci. 18:427-428(1993).

359. Prokaryotic molybdopterin oxidoreductases signatures (molybdopterin)

A number of different prokaryotic oxidoreductases that require and bind amolybdopterin cofactor have been shown [1,2,3] to share a number of regions of sequence similarity. These enzymes are: - Escherichia coli respiratory nitrate reductase (EC 1.7.99.4). This enzyme complex allows the bacteria to use nitrate as an electron acceptor during anaerobic growth. The enzyme is composed of three different chains: alpha, beta and gamma. The alpha chain

(gene narG) is the molybdopterin-binding subunit. *Escherichia coli* encodes for a second, closely related, nitrate reductase complex which also contains a molybdopterin-binding alpha chain (gene narZ). - *Escherichia coli* anaerobic dimethyl sulfoxide reductase (DMSO reductase). DMSO reductase is the terminal reductase during anaerobic growth on various sulfoxide and N-oxide compounds. DMSO reductase is composed of three chains: A, B and C. The A chain (gene dmsA) binds molybdopterin. - *Escherichia coli* biotin sulfoxide reductases (genes bisC and bisZ). This enzyme reduces a spontaneous oxidation product of biotin, BDS, back to biotin. It may serve as a scavenger, allowing the cell to use biotin sulfoxide as a biotin source. - *Methanobacterium formicicum* formate dehydrogenase (EC 1.2.1.2). The alpha chain (gene fdhA) of this dimeric enzyme binds a molybdopterin cofactor. - *Escherichia coli* formate dehydrogenases -H (gene fdhF), -N (gene fdnG) and -O (gene fdoG). These enzymes are responsible for the oxidation of formate to carbon dioxide. In addition to molybdopterin, the alpha (catalytic) subunit also contains an active site, selenocysteine. - *Wolinella succinogenes* polysulfide reductase chain. This enzyme is a component of the phosphorylative electron transport system with polysulfide as the terminal acceptor. It is composed of three chains: A, B and C. The A chain (gene psrA) binds molybdopterin. - *Salmonella typhimurium* thiosulfate reductase (gene phsA). - *Escherichia coli* trimethylamine-N-oxide reductase (EC 1.6.6.9) (gene torA) [4]. - Nitrate reductase (EC 1.7.99.4) from *Klebsiella pneumoniae* (gene nasA), *Alcaligenes eutrophus*, *Escherichia coli*, *Rhodobacter sphaeroides*, *Thiosphaera pantotropha* (gene napA), and *Synechococcus* PCC 7942 (gene narB). These proteins range from 715 amino acids (fdhF) to 1246 amino acids (narZ) in size. Three signature patterns for these enzymes were derived. The first is based on a conserved region in the N-terminal section and contains two cysteine residues perhaps involved in binding the molybdopterin cofactor. It should be noted that this region is not present in bisC. The second pattern is derived from a conserved region located in the central part of these enzymes.

Consensus pattern: [STAN SEQ ID NO:250]]-x-[CH]-x(2,3)-C-[STAG SEQ ID NO:20]]-[GSTVMF SEQ ID NO:376]]-x-C-x-[LIVMFYW SEQ ID NO:26]]-x-[LIVMA SEQ ID NO:30]]-x(3,4)-[DENQKHT SEQ ID NO:377]]-

Consensus pattern: [STA]-x-[STAC SEQ ID NO:204]](2)-x(2)-[STA]-D-[LIVMY SEQ ID NO:141]](2)-L-P-x-[STAC SEQ ID NO:204]](2)-x(2)-E-

Consensus pattern: A-x(3)-[GDT]-I-x-[DNQTK SEQ ID NO:378)]-x-[DEA]-x-[LIVM SEQ ID NO:4)]-x-[LIVMC SEQ ID NO:142)]-x-[NS]-x(2)-[GS]-x(5)-A-x-[LIVM SEQ ID NO:4)]-[ST]-

- 5 [1] Wootton J.C., Nicolson R.E., Cock J.M., Walters D.E., Burke J.F., Doyle W.A., Bray R.C. *Biochim. Biophys. Acta* 1057:157-185(1991).
 [2] Bilous P.T., Cole S.T., Anderson W.F., Weiner J.H. *Mol. Microbiol.* 2:785-795(1988).
 [3] Trieber C.A., Rothery R.A., Weiner J.H. *J. Biol. Chem.* 269:7103-7109(1994).
 [4] Mejean V., Lobbi-Nivol C., Lepelletier M., Giordano G., Chippaux M., Pascal M.-C.
 10 *Mol. Microbiol.* 11:1169-1179(1994).

360. Bacterial mutT domain signature

The bacterial mutT protein is involved in the GO system [1] responsible for removing an
 15 oxidatively damaged form of guanine (8-hydroxyguanine or 7,8-dihydro-8-oxoguanine) from DNA and the nucleotide pool. 8-oxo-dGTP is inserted opposite to dA and dC residues of template DNA with almost equal efficiency thus leading to A.T to G.C transversions. MutT specifically degrades 8-oxo-dGTP to the monophosphate with the concomitant release of pyrophosphate. MutT is a small protein of about 12 to 15 Kd. It has been shown [2,3] that a
 20 region of about 40 amino acid residues, which is found in the N-terminal part of mutT, can also be found in a variety of other prokaryotic, viral, and eukaryotic proteins. These proteins are:

- *Streptomyces pneumoniae* mutX.
- A mutT homolog from plasmid pSAM2 of *Streptomyces ambofaciens*.
- 25 - *Bartonella bacilliformis* invasion protein A (gene invA).
- *Escherichia coli* dATP pyrophosphohydrolase.
- Protein D250 from African swine fever viruses.
- Proteins D9 and D10 from a variety of poxviruses.
- Mammalian 7,8-dihydro-8-oxoguanine triphosphatase (EC 3.1.6.-) [4].
- 30 - Mammalian diadenosine 5',5'''-P1,P4-tetraphosphate asymmetrical hydrolase (Ap4Aase) (EC 3.6.1.17) [5], which cleaves A-5'-PPPP-5'A to yield AMP and ATP.

350

- A protein encoded on the antisense RNA of the basic fibroblast growth factor gene in higher vertebrates.
- Yeast protein YSA1.
- Escherichia coli hypothetical protein yfaO.
- 5 - Escherichia coli hypothetical protein ygdU and HI0901, the corresponding Haemophilus influenzae protein.
- Escherichia coli hypothetical protein yjaD and HI0432, the corresponding Haemophilus influenzae protein.
- Escherichia coli hypothetical protein yrfE.
- 10 - Bacillus subtilis hypothetical protein yqkG.
- Bacillus subtilis hypothetical protein yzgD.
- Yeast hypothetical protein YGL067w.

It is proposed [2] that the conserved domain could be involved in the active center of a family of pyrophosphate-releasing NTPases. As a signature pattern the core region of the domain was selected; it contains four conserved glutamate residues.

Consensus pattern: G-x(5)-E-x(4)-[STAGC SEQ ID NO:45)]-[LIVMAC SEQ ID NO:379)]-x-R-E-[LIVMFT SEQ ID NO:282)]-x-E-E-

- [1] Michaels M.L., Miller J.H. J. Bacteriol. 174:6321-6325(1992).
- [2] Koonin E.V. Nucleic Acids Res. 21:4847-4847(1993).
- [3] Mejean V., Salles C., Bullions M.J., Bessman M.J., Claverys J.-P. Mol. Microbiol. 11:323-330(1994).
- [4] Sakumi K., Furuichi M., Tsuzuki T., Kakuma T., Kawabata S., Maki H., Sekiguchi M. J. Biol. Chem. 268:23524-23530(1993).
- [5] Thorne N.M.H., Hankin S., Wilkinson M.C., Nunez C., Barraclough R., McLennan A.G. Biochem. J. 311:717-721(1995).

361. Myb DNA-binding domain repeat signatures

The retroviral oncogene v-myb, and its cellular counterpart c-myb, encode nuclear DNA-binding proteins that specifically recognize the sequence YAAC(G/T)G [1]. The myb family also includes the following proteins: - Drosophila D-myb [2]. - Vertebrate myb-like proteins A-myb and B-myb [3]. - Maize C1 protein, a trans-acting factor which controls the

expression of genes involved in anthocyanin biosynthesis. - Maize P protein [4], a trans-acting factor which regulates the biosynthetic pathway of a flavonoid-derived pigment in certain floral tissues. - Arabidopsis thaliana protein GL1 [5], required for the initiation of differentiation of leaf hair cells (trichomes). - A number of myb/c1-related proteins in maize and barley, whose roles are not yet known [4]. - Yeast BAS1 [7], a transcriptional activator for the HIS4 gene. - Yeast REB1 [8], which recognizes sites within both the enhancer and the promoter of rRNA transcription, as well as upstream of many genes transcribed by RNA polymerase II. - Fission yeast cdc5, a possible transcription factor whose activity is required for cell cycle progression and growth during G2. - Fission yeast myb1, which regulates telomere length and function. - Yeast hypothetical protein YMR213w. One of the most conserved regions in all of these proteins is a domain of 160 amino acids. It consists of three tandem repeats of 51 to 53 amino acids. In myb, this repeat region has been shown [9] to be involved in DNA-binding. The major part of the first repeat is missing in retroviral v-myb sequences and in plant myb-related proteins. Yeast REB1 differs from the other proteins in this family in having a single myb-like domain. As shown in the following schematic representation, two signature patterns for myb-like domains were developed; the first is located in the N-terminal section, the second spans the C-terminal extremity of the domain.

xxxxxxxxWxxxEDxxxxxxxxxxxxxxxxWxxIxxxxxxxxRxxxxxxxxWxxxx *****

*****' : Position of the patterns.

Consensus pattern: W-[ST]-x(2)-E-[DE]-x(2)-[LIV]-

Consensus pattern: W-x(2)-[LI]-[SAG]-x(4,5)-R-x(8)-[YW]-x(3)-[LIVM SEQ ID NO:4)]-

Note: this pattern detects the three copies of the domain in myb, d-myb, A-myb and B-myb; the second of the two complete copies of plant myb-related proteins, and the last two copies of yeast BAS1

[1] Biednkapp H., Borgmeyer U., Sippel A.E., Klempnauer K.-H. Nature 335:835-837(1988).

[2] Peters C.W.B., Sippel A.E., Vingron M., Klempnauer K.-H. EMBO J. 6:3085-3090(1987).

[3] Nomura N., Takahashi M., Matsui M., Ishii S., Date T., Sasamoto S., Ishizaki R. Nucleic Acids Res. 16:11075-11090(1988).

[4] Grotewold E., Athma P., Peterson T. Proc. Natl. Acad. Sci. U.S.A. 88:4587-4591(1991).

[5] Oppenheimer D.G., Herman P.L., Sivakumaran S., Esch J., Marks M.D. Cell 67:483-493(1991).

[6] Marocco A., Wissenbach M., Becker D., Paz-Ares J., Saedler H., Salamini F., Rohde W. Mol. Gen. Genet. 216:183-187(1989).

5 [7] Tice-Baldwin K., Fink G.R., Arndt K.T. Science 246:931-935(1989).

[8] Ju Q., Morrow B.E., Warner J.R. Mol. Cell. Biol. 10:5226-5234(1990).

[9] Klempnauer K.-H., Sippel A.E. EMBO J. 6:2719-2725(1987).

10 362. NAD-dependent glycerol-3-phosphate dehydrogenase signature

NAD-dependent glycerol-3-phosphate dehydrogenase (EC 1.1.1.8) (GPD) catalyzes the reversible reduction of dihydroxyacetone phosphate to glycerol-3- phosphate. It is a eukaryotic cytosolic homodimeric protein of about 40 Kd. As a signature pattern a glycine-rich region that is probably [1] involved in NAD-binding was selected.

15

Consensus pattern: G-[AT]-[LIVM SEQ ID NO:4]-K-[DN]-[LIVM SEQ ID NO:4])(2)-A-x-[GA]-x-G-[LIVMF SEQ ID NO:2)]-x- [DE]-G-[LIVM SEQ ID NO:4)]-x-[LIVMFYW SEQ ID NO:26)]-G-x-N-

20 [1] Otto J., Argos P., Rossmann M.G. Eur. J. Biochem. 109:325-330(1980).

363. Nucleosome assembly protein (NAP)

It is thought that NAPs may be involved in regulating gene expression as a result of histone accessibility [1].

25

[1] Rodriguez P, Munroe D, Prawitt D, Chu LL, Bric E, Kim J, Reid LH, Davies C, Nakagama H, Loebbert R, Winterpacht A, Petruzzi MJ, Higgins MJ, Nowak N, Evans G, Shows T, Weissman BE, Zabel B, Housman DE, Pelletier J, Genomics 1997;44:253-265. [2] Schnieders F, Dork T, Arnemann J, Vogel T, Werner M, Schmidtke J; Hum Mol Genet

30

364. NB-ARC domain

365. Nucleoside diphosphate kinases active site

5 Nucleoside diphosphate kinases (EC 2.7.4.6) (NDK) [1] are enzymes required for the synthesis of nucleoside triphosphates (NTP) other than ATP. They provide NTPs for nucleic acid synthesis, CTP for lipid synthesis, UTP for polysaccharide synthesis and GTP for protein elongation, signal transduction and microtubule polymerization. In eukaryotes, there seems to be a small family of NDK isozymes each of which acts in a different subcellular

10 compartment and/or has a distinct biological function. Eukaryotic NDK isozymes are hexamers of two highly related chains (A and B) [2]. By random association (A₆, A₅B...AB₅, B₆), these two kinds of chain form isoenzymes differing in their isoelectric point. NDK are proteins of 17 Kd that act via a ping-pong mechanism in which a histidine residue is phosphorylated, by transfer of the terminal phosphate group from ATP. In the presence of

15 magnesium, the phosphoenzyme can transfer its phosphate group to any NDP, to produce an NTP. NDK isozymes have been sequenced from prokaryotic and eukaryotic sources. It has also been shown [3] that the *Drosophila* awd (abnormal wing discs) protein, is a microtubule-associated NDK. Mammalian NDK is also known as metastasis inhibition factor nm23. The sequence of NDK has been highly conserved through evolution. There is a single histidine

20 residue conserved in all known NDK isozymes, which is involved in the catalytic mechanism [2]. Our signature pattern contains this residue.

Consensus pattern: N-x(2)-H-[GA]-S-D-[SA]-[LIVMPKNE SEQ ID NO:380)] [H is the putative active site residue]-

25 [1] Parks R., Agarwal R. (In) The Enzymes (3rd edition) 8:307-334(1973).

[2] Gilles A.-M., Presecan E., Vonica A., Lascu I. J. Biol. Chem. 266:8784-8789(1991).

[3] Biggs J., Hersperger E., Steeg P.S., Liotta L.A., Shearn A. Cell 63:933-940(1990).

366. Nitrite and sulfite reductases iron-sulfur/siroheme-binding site (NIR_SIR)

30 Nitrite reductases (NiR) [1] catalyze the reduction of nitrite into ammonium, the second step in the assimilation of nitrate. There are two types of NiR: the higher plant chloroplastic form

of NiR (EC 1.7.7.1) is a monomeric protein that uses reduced ferredoxin as the electron donor; while fungal and bacterial NiR (EC 1.6.6.4) are homodimeric proteins that uses NAD(P)H as the electron donor. Both forms of NiR contain a siroheme-Fe and iron-sulfur centers. Sulfite reductase (NADPH) (EC 1.8.1.2) (SIR) [2] is the bacterial enzyme that catalyzes the reduction of sulfite to sulfide. SIR is an oligomeric enzyme with a subunit composition of alpha(8)-beta(4), the alpha component is a flavoprotein (SIR-FP), while the beta component is a siroheme, iron-sulfurprotein (SIR-HP). Sulfite reductase (ferredoxin) (EC 1.8.7.1) [3] is a cyanobacterial and plant monomeric enzyme that also catalyzes the reduction of sulfite to sulfide. Anaerobic sulfite reductase (EC 1.8.1.-) (ASR) [4], a bacterial enzyme that catalyzes the NADH-dependent reduction of sulfite to sulfide. ASR is an oligomeric enzyme composed of three different subunits. The C component (geneasrC) seems to be a siroheme, iron-sulfur protein. These enzymes share a region of sequence similarity in their C-terminal half; this region which spans about 80 amino acids includes four conserved cysteine residues. Two of the Cys are grouped together at the beginning of the domain, and the two others are grouped in the middle of the domain. The cysteines are involved in the binding of the iron-sulfur center; the last one also binds the siroheme group [2]. A signature pattern from the region around the second cluster of cysteines was derived.

Consensus pattern: [STV]-G-C-x(3)-C-x(6)-[DE]-[LIVMF SEQ ID NO:2)]-[GAT]-[LIVMF SEQ ID NO:2)] [The two C's are iron-sulfur ligands]-

[1] Campbell W.H., Kinghorn J.R. Trends Biochem. Sci. 15:315-319(1990).

[2] Crane B.R., Siegel L.M., Getzoff E.D. Science 270:59-67(1995).

[3] Gisselmann G., Klausmeier P., Schwenn J.D. Biochim. Biophys. Acta 1144:102-106(1993).

[4] Huang C.J., Barrett E.L. J. Bacteriol. 173:1544-1553(1991).

367. (NMT) Myristoyl-CoA:protein N-myristoyltransferase signatures. Myristoyl-CoA: protein N-myristoyltransferase (EC 2.3.1.97) (Nmt) [1] is the enzyme responsible for transferring a myristate group on the N-terminal glycine of a number of cellular eukaryotic and viral proteins. Nmt is a monomeric protein of about 50 to 60 Kd whose sequence appears

to be well conserved. Two highly conserved regions have been developed as signature patterns. The first one is located in the central section, the second in the C-terminal part.

Consensus pattern: E-I-N-F-L-C-x-H-K-

5 Consensus pattern: K-F-G-x-G-D-G-

[1] Rudnick D.A., McWherter C.A., Gokel G.W., Gordon J.I. Adv. Enzymol. 67:375-430(1993).

10

368. ADP-glucose pyrophosphorylase signatures (NTP_transferase)

ADP-glucose pyrophosphorylase (glucose-1-phosphate adenylyltransferase) [1,2](EC 2.7.7.27) catalyzes a very important step in the biosynthesis of alpha 1,4-glucans (glycogen or starch) in bacteria and plants: synthesis of the activated glucosyl donor, ADP-glucose, from glucose-1-phosphate and ATP. ADP-glucose pyrophosphorylase is a tetrameric allosterically regulated enzyme. It is a homotetramer in bacteria while in plant chloroplasts and amyloplasts, it is a heterotetramer of two different, yet evolutionary related, subunits. There are a number of conserved regions in the sequence of bacterial and plant ADP-glucose pyrophosphorylase subunits. Three of these regions were selected as signature patterns. The first two are N-terminal and have been proposed to be part of the allosteric and/or substrate-binding sites in the Escherichia coli enzyme (gene glgC). The third pattern corresponds to a conserved region in the central part of the enzymes.

20

Consensus pattern: [AG]-G-G-x-G-[STK]-x-L-x(2)-L-[TA]-x(3)-A-x-P-A-[LV] -

25

Consensus pattern: W-[FY]-x-G-[ST]-A-[DNSH SEQ ID NO:381]-[AS]-[LIVMFYW SEQ ID NO:26])-

Consensus pattern: [APV]-[GS]-M-G-[LIVMN SEQ ID NO:382])-Y-[IVC]-[LIVMFY SEQ ID NO:18])-x(2)-[DENPHK SEQ ID NO:383)] -

30

[1] Nakata P.A., Greene T.W., Anderson J.M., Smith-White B.J., Okita T.W., Preiss J. Plant Mol. Biol. 17:1089-1093(1991).

[2] Preiss J., Ball K., Hutney J., Smith-White B.J., Li. L., Okitsa T.W. Pure Appl. Chem. 63:535-544(1991).

369. Sodium/hydrogen exchanger family

Na/H antiporters are key transporters in maintaining the pH of actively metabolizing cells. The molecular mechanisms of antiport are unclear.

These antiporters contain 10-12 transmembrane regions (M) at the amino-terminus and a large cytoplasmic region at the carboxyl terminus. The transmembrane regions M3-M12 share identity with other members of the family. The M6 and M7 regions are highly conserved. Thus, this is thought to be the region that is involved in the transport of sodium and hydrogen ions. The cytoplasmic region has little similarity throughout the family.

[1] Dibrov P, Flicgel L; FEBS Lett 1998;424:1-5. [2] Orlowski J, Grinstein S; J Biol Chem 1997;272:22373-22376. [3] Numata M, Petrecca K, Lake N, Orlowski J; J Biol Chem 1998;273:6951-6959.

370. Sodium:sulfate symporter family signature (Na_sulph_symp)

Integral membrane proteins that mediate the intake of a wide variety of molecules with the concomitant uptake of sodium ions (sodium symporters) can be grouped, on the basis of sequence and functional similarities into a number of distinct families. One of these families currently consists of the following proteins: - Mammalian sodium/sulfate cotransporter [1]. - Mammalian renal sodium/dicarboxylate cotransporter [2], which transports succinate and citrate. - Mammalian intestinal sodium/dicarboxylate cotransporter. - *Chlamydomonas reinhardtii* putative sulfur deprivation response regulator SAC1 [3]. - *Caenorhabditis elegans* hypothetical proteins B0285.6, F31F6.6, K08E5.2 and R107.1. - *Escherichia coli* hypothetical protein yfbS. - *Haemophilus influenzae* hypothetical protein HI0608. - *Synechocystis* strain PCC 6803 hypothetical protein sll0640. - *Methanococcus jannaschii* hypothetical protein MJ0672. These transporters are proteins of from 430 to 620 amino acids which are highly hydrophobic and which probably contain about 12 transmembrane regions. As a signature pattern, a conserved region was selected which is located in or near the penultimate transmembrane region.

Consensus pattern: [STACP SEQ ID NO:384]-S-x(2)-F-x(2)-P-[LIVM SEQ ID NO:4]-[GSA]-x(3)-N-x-[LIVM SEQ ID NO:4]-V-

- 5 [1] Markovich D., Forgo J., Stange G., Biber J., Murer H. Proc. Natl. Acad. Sci. U.S.A. 90:8073-8077(1993).
[2] Pajor A.M. Am. J. Physiol. 270:642-648(1996).
[3] Davies J.P., Yildiz F.H., Grossman A. EMBO J. 15:2150-2159(1996).

10

371. NifU-like domain

This is an alignment of the carboxy-terminal domain. This is the only common region between the NifU protein from nitrogen-fixing bacteria and rhodobacterial species. The biochemical function of NifU is unknown [1].

15

Ouzounis C, Bork P, Sander C, Trends Biochem Sci 1994;19:199-200.

372. Nitrilases / cyanide hydratase signatures

- 20 Nitrilases (EC 3.5.5.1) are enzymes that convert nitriles into their corresponding acids and ammonia. They are widespread in microbes as well as in plants where they convert indole-3-acetonitrile to the hormone indole-3-acetic acid. A conserved cysteine has been shown [1,2] to be essential for enzyme activity; it seems to be involved in a nucleophilic attack on the nitrile carbon atom. Cyanide hydratase (EC 4.2.1.66) converts HCN to formamide. In phytopathogenic fungi, it is used to avoid the toxic effect of cyanide released by wounded plants [3]. The sequence of cyanide hydrolase is evolutionary related to that of nitrilases.
- 25 Yeast hypothetical proteins YIL164c and YIL165c also belong to this family. As signature patterns for these enzymes, two conserved regions were selected. The first is located in the N-terminal section while the second, which contains the active site cysteine, is located in the central section.

30

Consensus pattern: G-x(2)-[LIVMFY SEQ ID NO:18])(2)-x-[IF]-x-E-x(2)-[LIVM SEQ ID NO:4])-x-G-Y-P-

Consensus pattern: G-[GAQ]-x(2)-C-[WA]-E-[NH]-x(2)-[PST]-[LIVMFYS SEQ ID NO:153])-x-[KR] [C is the active site residue]-

[1] Kobayashi M., Izui H., Nagasawa T., Yamada H. Proc. Natl. Acad. Sci. U.S.A. 90:247-251(1993).

[2] Kobayashi M., Komeda H., Yanaka N., Nagasawa T., Yamada H. J. Biol. Chem. 267:20746-20751(1992).

[3] Wang P., Vanetten H.D. Biochem. Biophys. Res. Commun. 187:1048-1054(1992).

373. NusB family

The NusB protein is involved in the regulation of rRNA biosynthesis by transcriptional antitermination.

Huenges M, Rolz C, Gschwind R, Peteranderl R, Berglechner F, Richter G, Bacher A, Kessler H, Gemmecker G, EMBO J 1998;17:4092-4100.

374. (Neur Chan) Neurotransmitter-gated ion-channels signature

Neurotransmitter-gated ion-channels [1,2,3,4] provide the molecular basis for rapid signal transmission at chemical synapses. They are post-synaptic oligomeric transmembrane complexes that transiently form a ionic channel upon the binding of a specific neurotransmitter. Presently, the sequence of subunits from five types of neurotransmitter-gated receptors are known: - The nicotinic acetylcholine receptor (AChR), an excitatory cation channel. In the motor endplates of vertebrates, it is composed of four different subunits (alpha, beta, gamma and delta or epsilon) with a molar stoichiometry of 2:1:1:1. In neurones, the AChR receptor is composed of two different types of subunits: alpha and non-alpha (also called beta). Nicotinic AChRs are also found in invertebrates. - The glycine receptor, an inhibitory chloride ion channel. The glycine receptor is a pentamer composed of two different subunits (alpha and beta). - The gamma-aminobutyric-acid (GABA) receptor, which is also an inhibitory chloride ion channel. The quaternary structure of the GABA receptor is complex; at least four classes of subunits are known to exist (alpha, beta, gamma, and delta) and there are many variants in each class (for example: six variants of the alpha class have already been sequenced). - The serotonin 5HT3 receptor. Serotonin is a biogenic hormone

that functions as a neurotransmitter, a hormone and a mitogen. There are seven major groups of serotonin receptors; six of these groups (5HT1, 5HT2, and 5HT4 to 5HT7) transduce extracellular signal by activating G proteins, while 5HT3 is a ligand-gated cation-specific ion channel which, when activated causes fast, depolarizing responses in neurons. - The glutamate receptor, an excitatory cation channel. Glutamate is the main excitatory neurotransmitter in the brain. At least three different types of glutamate receptors have been described and are named according to their selective agonists (kainate, N-methyl-D-aspartate (NMDA) and quisqualate). All known sequences of subunits from neurotransmitter-gated ion-channels are structurally related. They are composed of a large extracellular glycosylated N-terminal ligand-binding domain, followed by three hydrophobic transmembrane regions which form the ionic channel, followed by an intracellular region of variable length. A fourth hydrophobic region is found at the C-terminal of the sequence. The sequence of subunits from the AchR, GABA, 5HT3, and Gly receptors are clearly evolutionary related and share many regions of sequence similarities. These sequence similarities are either absent or very weak in the Glu receptors. In the N-terminal extracellular domain of AchR/GABA/5HT3/Gly receptors, there are two conserved cysteine residues, which, in AchR, have been shown to form a disulfide bond essential to the tertiary structure of the receptor. A number of amino acids between the two disulfide-bonded cysteines are also conserved. Therefore this region was used as a signature pattern for this subclass of proteins.

Consensus pattern: C-x-[LIVMFQ SEQ ID NO:385)]-x-[LIVMF SEQ ID NO:2)]-x(2)-[FY]-P-x-D-x(3)-C [The two C's are linked by a disulfide bond]-

[1] Stroud R.M., McCarthy M.P., Shuster M. Biochemistry 29:11009-11023(1990).

[2] Betz H. Neuron 5:383-392(1990).

[3] Dingledine R., Myers S.J., Nicholas R.A. FASEB J. 4:2632-2645(1990).

[4] Barnard E.A. Trends Biochem. Sci. 17:368-374(1992).

375. Orotidine 5'-phosphate decarboxylase active site

Orotidine 5'-phosphate decarboxylase (EC 4.1.1.23) (OMPdecase) [1,2] catalyzes the last step in the de novo biosynthesis of pyrimidines, the decarboxylation of OMP into UMP. In higher eukaryotes OMPdecase is part, with orotatephosphoribosyltransferase, of a bifunctional

360

enzyme, while the prokaryotic and fungal OMPdecases are monofunctional protein. Some parts of the sequence of OMPdecase are well conserved across species. The best conserved region is located in the N-terminal half of OMPdecases and is centered around a lysine residue which is essential for the catalytic function of the enzyme. This region has been developed as a signature pattern.

Consensus pattern: [LIVMFTA SEQ ID NO:386]-[LIVMF SEQ ID NO:2)]-x-D-x-K-x(2)-D-I-[GP]-x-T-[LIVMTA SEQ ID NO:311)] [K is the active site residue]-

[1] Jacquet M., Guilbaud R., Garreau H. Mol. Gen. Genet. 211:441-445(1988).
[2] Kimsey H.H., Kaiser D. J. Biol. Chem. 267:819-824(1992).

376. ATP synthase delta (OSCP) subunit signature

ATP synthase (proton-translocating ATPase) (EC 3.6.1.34) [1,2] is a component of the cytoplasmic membrane of eubacteria, the inner membrane of mitochondria, and the thylakoid membrane of chloroplasts. The ATPase complex is composed of an oligomeric transmembrane sector, called CF(0), which acts as a proton channel, and a catalytic core, termed coupling factor CF(1).

One of the subunits of the ATPase complex, known as subunit delta in bacteria and chloroplasts or the Oligomycin Sensitivity Conferral Protein (OSCP) in mitochondria, seems to be part of the stalk that links CF(0) to CF(1). It either transmits conformational changes from CF(0) into CF(1) or is involved in proton conduction [3].

The different delta/OSCP subunits are proteins of approximately 200 amino-acid residues - once the transit peptide has been removed in the chloroplast and mitochondrial forms - which show only moderate sequence homology. The signature pattern used to detect ATPase delta/OSCP subunits is based on a conserved region in the C-terminal section of these proteins.

Consensus pattern: [LIVM SEQ ID NO:4)]-x-[LIVMFYT SEQ ID NO:143)]-x(3)-[LIVMT SEQ ID NO:1)]-[DENQK SEQ ID NO:387)]-x(2)-[LIVM SEQ ID NO:4)]-x-[GSA]-G-

[LIVMFYGA SEQ ID NO:388)]-x-[LIVM SEQ ID NO:4)]-[KRHENQ SEQ ID NO:389)]-x-[GSEN SEQ ID NO:390)]

[1] Futai M., Noumi T., Maeda M. Annu. Rev. Biochem. 58:111-136(1989).

5 [2] Senior A.E. Physiol. Rev. 68:177-231(1988).

[3] Engelbrecht S., Junge W. Biochim. Biophys. Acta 1015:379-390(1990).

377. Aspartate and ornithine carbamoyltransferases signature

10 Aspartate carbamoyltransferase (EC 2.1.3.2) (ATCase) catalyzes the conversion of aspartate and carbamoyl phosphate to carbamoylaspartate, the second step in the de novo biosynthesis of pyrimidine nucleotides [1]. In prokaryotes ATCase consists of two subunits: a catalytic chain (gene pyrB) and a regulatory chain (gene pyrI), while in eukaryotes it is a domain in a multi-
15 functional enzyme (called URA2 in yeast, rudimentary in Drosophila, and CAD in mammals [2]) that also catalyzes other steps of the biosynthesis of pyrimidines.

Ornithine carbamoyltransferase (EC 2.1.3.3) (OTCase) catalyzes the conversion of ornithine and carbamoyl phosphate to citrulline. In mammals this enzyme
20 participates in the urea cycle [3] and is located in the mitochondrial matrix. In prokaryotes and eukaryotic microorganisms it is involved in the biosynthesis of arginine. In some bacterial species it is also involved in the degradation of arginine [4] (the arginine deaminase pathway).

It has been shown [5] that these two enzymes are evolutionary related. The
25 predicted secondary structure of both enzymes are similar and there are some regions of sequence similarities. One of these regions includes three residues which have been shown, by crystallographic studies [6], to be implicated in binding the phosphoryl group of carbamoyl phosphate.

This region was selected as a signature for these enzymes.

30

Consensus pattern: F-x-[EK]-x-S-[GT]-R-T[S, R, and the 2nd T bind carbamoyl phosphate]

-Note: the residue in position 3 of the pattern allows to distinguish between an ATCase (Glu) and an OTCase (Lys).

- [1] Lerner C.G., Switzer R.L. J. Biol. Chem. 261:11156-11165(1986).
- [2] Davidson J.N., Chen K.C., Jamison R.S., Musmanno L.A., Kern C.B. BioEssays 15:157-164(1993).
- 5 [3] Takiguchi M., Matsubasa T., Amaya Y., Mori M. BioEssays 10:163-166(1989).
- [4] Baur H., Stalon V., Falmagne P., Luethi E., Haas D. Eur. J. Biochem. 166:111-117(1987).
- [5] Houghton J.E., Bencini D.A., O'Donovan G.A., Wild J.R. Proc. Natl. Acad. Sci. U.S.A. 81:4864-4868(1981).
- 10 [6] Ke H.-M., Honzatko R.B., Lipscomb W.N. Proc. Natl. Acad. Sci. U.S.A. 81:4037-4040(1984).

378. Oleosins signature

- 15 Oleosins [1] are the proteinaceous components of plants' lipid storage bodies called oil bodies. Oil bodies are small droplets (0.2 to 1.5 μ m in diameter) containing mostly triacylglycerol that are surrounded by a phospholipid/oleosin annulus. Oleosins may have a structural role in stabilizing the lipid body during dessication of the seed, by preventing coalescence of the oil.
- 20 They may also provide recognition signals for specific lipase anchorage in lipolysis during seedling growth. Oleosins are found in the monolayer lipid/water interface of oil bodies and probably interact with both the lipid and phospholipid moieties.
- Oleosins are proteins of 16 Kd to 24 Kd and are composed of three domains: an
- 25 N-terminal hydrophilic region of variable length (from 30 to 60 residues); a central hydrophobic domain of about 70 residues and a C-terminal amphipathic region of variable length (from 60 to 100 residues). The central hydrophobic domain is proposed to be made up of beta-strand structure and to interact with the lipids [2]. It is the only domain whose sequence is conserved and therefore
- 30 a section from that domain was selected as a signature pattern.

Consensus pattern: [AG]-[ST]-x(2)-[AG]-x(2)-[LIVM SEQ ID NO:4)]-[SAD]-T-P-[LIVMF SEQ ID NO:2)](4)-F-S-P-[LIVM SEQ ID NO:4)](3)-P-A

- [1] Murphy D.J., Keen J.N., O'Sullivan J.N., Au D.M.Y., Edwards E.-W., Jackson P.J., Cummins I., Gibbons T., Shaw C.H., Ryan A.J. *Biochim. Biophys. Acta* 1088:86-94(1991).
[2] Tzen J.T.C., Lie G.C., Huang A.H.C. *J. Biol. Chem.* 267:15626-15634(1992).

5

379. (Orbi VP5) Orbivirus outer capsid protein VP5

This paper shows the location of the different capsid proteins
and their relation to each other.

10

- [1] Schoehn G, Moss SR, Nuttall PA, Hewat EA; *Virology* 1997;235:191-200.

15 380. Orn/DAP/Arg decarboxylases family 2 signatures

Pyridoxal-dependent decarboxylases acting on ornithine, lysine, arginine and related substrates can be classified into two different families on the basis of sequence similarities [1,2,3]. The second family consists of:

- Eukaryotic ornithine decarboxylase (EC 4.1.1.17) (ODC). ODC catalyzes the transformation of ornithine into putrescine.
- Prokaryotic diaminopimelic acid decarboxylase (EC 4.1.1.20) (DAPDC). DAPDC catalyzes the conversion of diaminopimelic acid into lysine; the last step in the biosynthesis of lysine.
- *Pseudomonas syringae* pv. *tabaci* protein tabA. tabA is probably involved in the biosynthesis of tabtoxin and is highly similar to DAPDC.
- Bacterial and plant biosynthetic arginine decarboxylase (EC 4.1.1.19) (ADC). ADC catalyzes the transformation of arginine into agmatine, the first step in the biosynthesis of putrescine from arginine.

20

The above proteins, while most probably evolutionary related, do not share extensive regions of sequence similarities. Two of the conserved regions were selected as signature patterns. The first pattern contains a conserved lysine residue which is known, in mouse ODC [4], to be the site of attachment of the pyridoxal-phosphate group. The second pattern contains a stretch of three

30

consecutive glycine residues and has been proposed to be part of a substrate-binding region [5].

These enzymes are collectively known as group IV decarboxylases [3].

- 5 Consensus pattern: [FY]-[PA]-x-K-[SACV SEQ ID NO:391)]-[NHCLFW SEQ ID NO:392)]-x(4)-[LIVMF SEQ ID NO:2)]-[LIVMTA SEQ ID NO:311)]-x(2)- [LIVMA SEQ ID NO:30)]-x(3)-[GTE] [K is the pyridoxal-P attachment site]
Consensus pattern: [GS]-x(2,6)-[LIVMSCP SEQ ID NO:393)]-x(2)-[LIVMF SEQ ID NO:2)]-[DNS]-[LIVMCA SEQ ID NO:149)]-G-G-G-[LIVMFY SEQ ID NO:18)]-
10 [GSTPCEQ SEQ ID NO:394)]

[1] Bairoch A. Unpublished observations (1993).

[2] Martin C., Cami B., Yeh P., Stragier P., Parsot C., Patte J.-C. Mol. Biol. Evol. 5:549-559(1988).

- 15 [3] Sandmeier E., Hale T.I., Christen P. Eur. J. Biochem. 221:997-1002(1994).

[4] Poulin R., Lu L., Ackermann B., Bey P., Pegg A.E. J. Biol. Chem. 267:150-158(1992).

[5] Moore R.C., Boyle S.M. J. Bacteriol. 172:4631-4640(1990).

20 381. Osteopontin signature

Osteopontin is an acidic phosphorylated glycoprotein of about 40 Kd which is abundant in the mineral matrix of bones and which binds tightly to hydroxyapatite [1,2,3]. It is suggested that osteopontin might function as a cell attachment factor and could play a key role in the adhesion of

- 25 osteoclasts to the mineral matrix of bone.

Osteopontin-K is a kidney protein which is highly similar to osteopontin and probably also involved in cell-adhesion.

As a signature pattern a highly conserved region located at the N-terminal extremity of the mature protein was selected.

30

Consensus pattern: [KQ]-x-[TA]-x(2)-[GA]-S-S-E-E-K

[1] Butler W.T. Connect. Tissue Res. 23:123-36(1989).

[2] Gorski J.P. Calcif. Tissue Int. 50:391-396(1992).

[3] Denhardt D.T., Guo X. FASEB J. 7:1475-1482(1993).

5 382. Oxysterol-binding protein family signature

A number of eukaryotic proteins that seem to be involved with sterol synthesis and/or its regulation have been found [1] to be evolutionary related:

- Mammalian oxysterol-binding protein (OSBP). A protein of about 800 amino-acid residues that binds a variety of oxysterols: oxygenated derivatives of cholesterol. OSBP seems to play a complex role in the regulation of sterol metabolism.
- Yeast proteins HES1 and KES1; highly related proteins of 434 residues that seem to play a role in ergosterol synthesis.
- Yeast OSH1, a protein of 859 residues that also plays a role in ergosterol synthesis.
- Yeast hypothetical protein YHR001w (437 residues).
- Yeast hypothetical protein YHR073w (996 residues).
- Yeast hypothetical protein YKR003w (448 residues).

15 All these proteins contain a moderately conserved domain of about 250 residues located in the C-terminal half of OBSP, OSH1 and YHR073w and in the central section of the other proteins. As a signature pattern, the best conserved part was selected of this domain, a region that contains a conserved pentapeptide.

Consensus pattern: E-[KQ]-x-S-H-[HR]-P-P-x-[STACF SEQ ID NO:395)]-A

25 [1] Jiang B., Brown J.L., Sheraton J., Fortin N., Bussey H. Yeast 10:341-353(1994).

383. FMN oxidoreductase

384. Oxidoreductase FAD/NAD-binding domain

Number of members: 250

[1]

Medline: 92084635

The sequence of squash NADH:nitrate reductase and its relationship to the sequences of other flavoprotein oxidoreductases. A family of flavoprotein pyridine nucleotide cytochrome reductases.

Hyde GE, Crawford NM, Campbell W;

J Biol Chem 1991;266:23542-23547.

[2]Medline: 95111952

Crystal structure of the FAD-containing fragment of corn nitrate reductase at 2.5 Å resolution: relationship to other flavoprotein reductases.

Lu G, Campbell WH, Schneider G, Lindqvist Y;

Structure 1994;2:809-821.

385. (oxidored molyb) Eukaryotic molybdopterin oxidoreductases signature
A number of different eukaryotic oxidoreductases that require and bind a molybdopterin cofactor have been shown [1] to share a few regions of sequence similarity. These enzymes are:

- Xanthine dehydrogenase (EC 1.1.1.204), which catalyzes the oxidation of xanthine to uric acid with the concomitant reduction of NAD. Structurally, this enzyme of about 1300 amino acids consists of at least three distinct domains: an N-terminal 2Fe-2S ferredoxin-like iron-sulfur binding domain (see <PDOC00175>), a central FAD/NAD-binding domain and a C-terminal Molybdopterin domain.
- Aldehyde oxidase (EC 1.2.3.1), which catalyzes the oxidation of aldehydes into acids. Aldehyde oxidase is highly similar to xanthine dehydrogenase in its sequence and domain structure.
- Nitrate reductase (EC 1.6.6.1), which catalyzes the reduction of nitrate to nitrite. Structurally, this enzyme of about 900 amino acids consists of an N-terminal Molybdopterin domain, a central cytochrome b5-type heme-binding domain (see <PDOC00170>) and a C-terminal FAD/NAD-binding cytochrome

reductase domain.

- Sulfite oxidase (EC 1.8.3.1), which catalyzes the oxidation of sulfite to sulfate. Structurally, this enzyme of about 460 amino acids consists of an N-terminal cytochrome b5-binding domain followed by a Mo-pterin domain.

5 There are a few conserved regions in the sequence of the molybdopterin-binding domain of these enzymes. The pattern used to detect these proteins is based on one of them. It contains a cysteine residue which could be involved in binding the molybdopterin cofactor.

10 Consensus pattern: [GA]-x(3)-[KRNQHT SEQ ID NO:396)]-x(11,14)-[LIVMFYWS SEQ ID NO:301)]-x(8)-[LIVMF SEQ ID NO:2)]-x-C-x(2)-[DEN]-R-x(2)-[DE]

[1] Wootton J.C., Nicolson R.E., Cock J.M., Walters D.E., Burke J.F., Doyle W.A., Bray R.C. Biochim. Biophys. Acta 1057:157-185(1991).

15

386. (Oxidored q1) NADH-Ubiquinone/plastoquinone (complex I), various chains

This family is part of complex I which catalyses the transfer of two electrons from NADH to ubiquinone in a reaction that is associated with proton translocation across the membrane. Number of members: 1824

20

[1]

Medline: 93110040

The NADH:ubiquinone oxidoreductase (complex I) of respiratory chains. Walker JE;

25

Q Rev Biophys 1992;25:253-324.

387. (oxidored q3) NADH-ubiquinone/plastoquinone oxidoreductase chain 6. 179 members.

30

388. (oxidored q5) NADH-ubiquinone oxidoreductase chain 4, amino terminus

[1] Walker JE ; Q Rev Biophys 1992;25:253-324.

389. (oxidored q6) Respiratory-chain NADH dehydrogenase 20 Kd subunit signature
 Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complex
 5 I or NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex
 located in the inner mitochondrial membrane which also seems to exist in
 the chloroplast and in cyanobacteria (as a NADH-plastoquinone oxidoreductase).
 Among the 25 to 30 polypeptide subunits of this bioenergetic enzyme complex
 there is one with a molecular weight of 20 Kd (in mammals) [3], which is a
 10 component of the iron-sulfur (IP) fragment of the enzyme. It seems to bind a
 4Fe-4S iron-sulfur cluster. The 20 Kd subunit has been found to be:

- Nuclear encoded, as a precursor form with a transit peptide in mammals, and
 in *Neurospora crassa*. - Mitochondrial encoded in *Paramecium* (gene *psbG*).
- Chloroplast encoded in various higher plants (gene *ndhK* or *psbG*).

15 The 20 Kd subunit is highly similar to [4]:

- *Synechocystis* strain PCC 6803 proteins *psbG1* and *psbG2*.
- Subunit B of *Escherichia coli* NADH-ubiquinone oxidoreductase (gene *nuoB*).
- Subunit NQO6 of *Paracoccus denitrificans* NADH-ubiquinone oxidoreductase.
- Subunit 7 of *Escherichia coli* formate hydrogenlyase (gene *hycG*).
- 20 - Subunit I of *Escherichia coli* hydrogenase-4 (gene *hyfI*).

As as signature pattern a highly conserved region was selected, located in the
 central section of this subunit and which contains a conserved cysteine that
 is probably involved in the binding of the 4Fe-4S center.

25 Consensus pattern: [GN]-x-D-[KRST SEQ ID NO:397)]-[LIVMF SEQ ID NO:2)](2)-P-[IV]-
 D-[LIVMFYW SEQ ID NO:26)](2)-x-P-x-C-P-[PT] [The C is a putative 4Fe-4S ligand]
 [1] Ragan C.I. Curr. Top. Bioenerg. 15:1-36(1987).

[2] Weiss H., Friedrich T., Hofhaus G., Preis D. Eur. J. Biochem. 197:563-576(1991).

[3] Arizmendi J.M., Runswick M.J., Skehel J.M., Walker J.E. FEBS Lett. 301:237-
 30 242(1992).

[4] Weidner U., Geier S., Ptock A., Friedrich T., Leif H., Weiss H. J. Mol. Biol. 233:109-
 122(1993).

390. p53 tumor antigen signature

The p53 tumor antigen [1 to 5, E1,E2] is a protein found in increased amounts in a wide variety of transformed cells. It is also detectable in many proliferating nontransformed cells, but it is undetectable or present at low levels in resting cells. It is frequently mutated or inactivated in many types of cancer. p53 seems to act as a tumor suppressor in some, but probably not all, tumor types. p53 is probably involved in cell cycle regulation, and may be a trans-activator that acts to negatively regulate cellular division by controlling a set of genes required for this process.

p53 is a phosphoprotein of about 390 amino acids which can be subdivided into four domains: a highly charged acidic region of about 75 to 80 residues, a hydrophobic proline-rich domain (position 80 to 150), a central region (from 150 to about 300), and a highly basic C-terminal region. The sequence of p53 is well conserved in vertebrate species; attempts to identify p53 in other eukaryotic phylum has so far been unsuccessful.

As a signature pattern for p53 a perfectly conserved stretch of 13 residues located in the central region of the protein was selected. This region, known as domain IV in [3], is involved (along with an adjacent region) in the binding of the large T antigen of SV40. In man this region is the focus of a variety of point mutations in cancerous tumors.

Consensus pattern: M-C-N-S-S-C-M-G-G-M-N-R-R

[1] Levine A.J., Momand J., Finlay C.A. Nature 351:453-456(1991).

[2] Levine A.J., Momand J. Biochim. Biophys. Acta 1032:119-136(1990).

[3] Soussi T., Caron De Fromentel C., May P. Oncogene 5:945-952(1990).

[4] Lane D.P., Benchimol S. Genes Dev. 4:1-8(1990).

[5] Ulrich S.J., Anderson C.W., Mercer W.E., Appella E. J. Biol. Chem. 267:15259-15262(1992).

391. (P5CR) Delta 1-pyrroline-5-carboxylate reductase signature

Delta 1-pyrroline-5-carboxylate reductase (P5CR) (EC 1.5.1.2) [1,2] is the

370

enzyme that catalyzes the terminal step in the biosynthesis of proline from glutamate, the NAD(P) dependent oxidation of 1-pyrroline-5-carboxylate into proline.

The sequences of P5CR from eubacteria (gene proC), archaeobacteria and eukaryotes show only a moderate level of overall similarity. As a signature pattern, the best conserved region located in the C-terminal section of P5CR was selected.

Consensus pattern: [PALF SEQ ID NO:398)]-x(2,3)-[LIV]-x(3)-[LIVM SEQ ID NO:4)]-[STAC SEQ ID NO:204)]-[STV]-x-[GAN]-G-x-T-x(2)-[AG]-[LIV]-x(2)-[LMF]-[DENQK SEQ ID NO:387)]

[1] Delauney A.J., Verma D.P. Mol. Gen. Genet. 221:299-305(1990).

[2] Savioz A., Jeenes D.J., Kocher H.P., Haas D. Gene 86:107-111(1990).

392. Poly-adenylate binding protein, unique domain.

393. (PAL) Phenylalanine and histidine ammonia-lyases active site

Phenylalanine ammonia-lyase (EC 4.3.1.5) (PAL) is a key enzyme of plant and fungi phenylpropanoid metabolism which is involved in the biosynthesis of a wide variety of secondary metabolites such as flavanoids, furanocoumarin phytoalexins and cell wall components. These compounds have many important roles in plants during normal growth and in responses to environmental stress. PAL catalyzes the removal of an ammonia group from phenylalanine to form trans-cinnamate.

Histidine ammonia-lyase (EC 4.3.1.3) (histidase) catalyzes the first step in histidine degradation, the removal of an ammonia group from histidine to produce urocanic acid.

The two types of enzymes are functionally and structurally related [1]. They are the only enzymes which are known to have the modified amino acid dehydro-alanine (DHA) in their active site. A serine residue has been shown [2,3,4] to

be the precursor of this essential electrophilic moiety. The region around this active site residue is well conserved and can be used as a signature pattern.

- 5 Consensus pattern: G-[STG]-[LIVM SEQ ID NO:4)]-[STG]-[AC]-S-G-[DH]-L-x-P-L-[SA]-x(2)-[SA] [S is the active site residue]

- [1] Taylor R.G., Lambert M.A., Sexsmith E., Sadler S.J., Ray P.N., Mahuran D.J., McInnes R.R. J. Biol. Chem. 265:18192-18199(1990).
- 10 [2] Langer M., Reck G., Reed J., Retey J. Biochemistry 33:6462-6467(1994).
- [3] Schuster B., Retey J. FEBS Lett. 349:252-254(1994).
- [4] Taylor R.G., McInnes R.R. J. Biol. Chem. 269:27473-27477(1994).

15 394. PAS domain

-!- CAUTION. This family does not currently match all known examples of PAS domains.

PAS motifs appear in archaea, eubacteria and eukarya. Probably the most surprising identification of a PAS domain was that in

- 20 EAG-like K⁺-channels[1,3].

Number of members: 308

[1]

Medline: 97446881

PAS domain S-boxes in archaea, bacteria and sensors for oxygen and redox.

25 Zhulin IB, Taylor BL, Dixon R;

Trends Biochem Sci 1997;22:331-333.

[2]Medline: 95275818

1.4 A structure of photoactive yellow protein, a cytosolic

- 30 photoreceptor: unusual fold, active site, and chromophore.

Borgstahl GE, Williams DR, Getzoff ED;

Biochemistry 1995;34:6278-6287.

[3]Medline: 98044337

PAS: a multifunctional domain family comes to light.

Ponting CP, Aravind L;

Curr Biol 1997;7:674-677.

5

395. (PBP) Phosphatidylethanolamine-binding protein family signature

Mammalian phosphatidylethanolamine-binding protein (also known as basic cytosolic 21 Kd protein) is a 186 residue protein found in a variety of tissues [1]. It binds hydrophobic ligands, such as phosphatidylethanolamine, but also seems [2] to bind nucleotides such as GTP and FMN, it is suggested that it could act in membrane remodeling during growth and maturation. This protein belongs to a family that also includes:

10

- Drosophila antennal protein A5, a putative odorant-binding protein.
- Onchocerca volvulus antigen Ov-16 and the related proteins D1, D2 and D3.
- 15 - Plasmodium falciparum putative phosphatidylethanolamine-binding protein.
- Toxocara canis secreted antigen TES-26. This larval protein has been shown to bind phosphatidylethanolamine.
- Yeast protein DKA1 (also known as NSP1 or TFS1). The function of this protein is not very clear.
- Yeast hypothetical protein YLR179C.
- 20 - Caenorhabditis elegans hypothetical protein F40A3.3.

As a signature pattern, the best conserved region was selected which is located in the end of the first third of the sequence of these proteins.

25

Consensus pattern: [FYL]-x-[LV]-[LIVF SEQ ID NO:127)]-x-[TIV]-[DC]-P-D-x-P-[SN]-x(10)-H

[1] Seddiqui N., Bollengier F., Alliel P.M., Perin J.P., Bonnet F., Bucquoy S., Jolles P., Schoentgen F. J. Mol. Evol. 39:655-660(1994).

[2] Schoentgen F., Jolles P. FEBS Lett. 369:22-6(1995).

30

396. PCI domain

This domain has also been called the PINT motif (Proteasome,

Int-6, Nip-1 and TRIP-15) [1].

Number of members: 49

[1]

Medline: 98308842

- 5 The PCI domain: a common theme in three multiprotein complexes.

Hofmann K, Bucher P;

Trends Biochem Sci 1998;23:204-205.

[2]Medline: 98266368

- 10 Homologues of 26S proteasome subunits are regulators of transcription and translation.

Aravind L, Ponting CP;

Protein Sci 1998;7:1250-1254.

15

397. (PCMT) Protein-L-isoaspartate (D-aspartate) O-methyltransferase signature. Protein-L-isoaspartate (D-aspartate) O-methyltransferase (EC 2.1.1.77) (PCMT)[1] (which is also known as L-isoaspartyl protein carboxyl methyltransferase) is an enzyme that catalyzes the transfer of a methyl group from S-adenosylmethionine to the free carboxyl groups of D-aspartyl or L-isoaspartyl residues in a variety of peptides and proteins. The enzyme does not act on normal L-aspartyl residues L-isoaspartyl and D-aspartyl are the products of the spontaneous de amidation and/or isomerization of normal L-aspartyl and L-asparaginyl residues in proteins. PCMT plays a role in the repair and/or degradation of these damaged proteins; the enzymatic methyl esterification of the abnormal residues can lead to their conversion to normal L-aspartyl residues. PCMT is a well-conserved and widely distributed cytosolic protein of about 24Kd. As a signature pattern, a conserved region in the central part of this enzyme has been developed.

25

Consensus pattern: [GSA]-D-G-x(2)-G-[FYWV SEQ ID NO:399)]-x(3)-[AS]-P-[FY]-[DN]-x-I -

30

[1] Kagan R.M., McFadden H.J., McFadden P.N., O'Connor C., Clarke S. Comp. Biochem. Physiol. 117b:379-385(1997).

398. (PCNA) Proliferating cell nuclear antigen signatures

Proliferating cell nuclear antigen (PCNA) [1,2] is a protein involved in DNA replication by acting as a cofactor for DNA polymerase delta, the polymerase responsible for leading strand DNA replication.

A similar protein exists in yeast (gene POL30) [3] and is associated with polymerase III, the yeast analog of polymerase delta. In baculoviruses the ETL protein has been shown [4] to be highly related to PCNA and is probably associated with the viral encoded DNA polymerase. An homolog of PCNA is also found in archebacteria.

As signatures for this family of proteins, two conserved regions were selected located in the N-terminal section. The second one has been proposed to bind DNA.

Consensus pattern: [GA]-[LIVMF SEQ ID NO:2)]-x-[LIVMA SEQ ID NO:30)]-x-[SAV]-[LIVM SEQ ID NO:4)]-D-x-[NSAE SEQ ID NO:400)]-[HKR]-[VI]-x-[LY]-[VGA]-x-[LIVM SEQ ID NO:4)]-x-[LIVM SEQ ID NO:4)]-x(4)-F

-Consensus pattern: [RKA]-C-[DE]-[RH]-x(3)-[LIVMF SEQ ID NO:2)]-x(3)-[LIVM SEQ ID NO:4)]-x-[SGAN SEQ ID NO:401)]-[LIVMF SEQ ID NO:2)]-x-K-[LIVMF SEQ ID NO:2)](2)

[1] Bravo R., Frank R., Blundell P.A., McDonald-Bravo H. Nature 326:515-517(1987).

[2] Suzuka I., Hata S., Matsuoka M., Kosugi S., Hashimoto J. Eur. J. Biochem. 195:571-

575(1991).[3] Bauer G.A., Burgess P.M.J. Nucleic Acids Res. 18:261-265(1990).

[4] O'Reilly D.R., Crawford A.M., Miller L.K. Nature 337:606-606(1989).

399. (PDT) Prephenate dehydratase signatures

Prephenate dehydratase (EC 4.2.1.51) (PDT) catalyzes the decarboxylation of prephenate into phenylpyruvate. In microorganisms PDT is involved in the terminal pathway of the biosynthesis of phenylalanine. In some bacteria such as Escherichia coli PDT is part of a bifunctional enzyme (P-protein) that also

catalyzes the transformation of chorismate into prephenate (chorismate mutase) while in other bacteria it is a monofunctional enzyme. The sequence of monofunctional PDT align well with the C-terminal part of that of P-proteins [1].

5 As signature patterns for PDT two conserved regions were selected. The first region contains a conserved threonine which has been said to be essential for the activity of the enzyme in *E. coli*. The second region includes a conserved glutamate. Both regions are in the C-terminal part of PDT.

10 Consensus pattern: [FY]-x-[LIVM SEQ ID NO:4)]-x(2)-[LIVM SEQ ID NO:4)]-x(5)-[DN]-x(5)-T-R-F-[LIVMW SEQ ID NO:235)]-x-[LIVM SEQ ID NO:4)]

[1] Fischer R.S., Zhao G., Jensen R.A. J. Gen. Microbiol. 137:1293-1301(1991).

15

400. PDZ domain (Also known as DHR or GLGF).

PDZ domains are found in diverse signaling proteins.

[1] Ponting CP, Phillips C, Davies KE, Blake DJ

Bioessays 1997;19:469-479. [2] Doyle DA, Lee A, Lewis J, Kim E, Sheng M, MacKinnon R;

20 Cell. 1996;85:1067-1076. [3] Ponting CP; Protein Sci 1997;6:464-468.

401. (PPDK_N_term) PEP-utilizing enzymes signatures

A number of enzymes that catalyze the transfer of a phosphoryl group from phosphoenolpyruvate (PEP) via a phospho-histidine intermediate have been shown to be structurally related [1,2,3,4]. These enzymes are:

- Pyruvate, orthophosphate dikinase (EC 2.7.9.1) (PPDK). PPDK catalyzes the reversible phosphorylation of pyruvate and phosphate by ATP to PEP and diphosphate. In plants PPDK function in the direction of the formation of

30 PEP, which is the primary acceptor of carbon dioxide in C4 and crassulacean acid metabolism plants. In some bacteria, such as *Bacteroides symbiosus*, PPDK functions in the direction of ATP synthesis.

- Phosphoenolpyruvate synthase (EC 2.7.9.2) (pyruvate, water dikinase). This

enzyme catalyzes the reversible phosphorylation of pyruvate by ATP to form PEP, AMP and phosphate, an essential step in gluconeogenesis when pyruvate and lactate are used as a carbon source.

- Phosphoenolpyruvate-protein phosphotransferase (EC 2.7.3.9). This is the first enzyme of the phosphoenolpyruvate-dependent sugar phosphotransferase system (PTS), a major carbohydrate transport system in bacteria. The PTS catalyzes the phosphorylation of incoming sugar substrates concomitant with their translocation across the cell membrane. The general mechanism of the PTS is the following: a phosphoryl group from PEP is transferred to enzyme-I (EI) of PTS which in turn transfers it to a phosphoryl carrier protein (HPr). Phospho-HPr then transfers the phosphoryl group to a sugar-specific permease.

All these enzymes share the same catalytic mechanism: they bind PEP and transfer the phosphoryl group from it to a histidine residue. The sequence around that residue is highly conserved and can be used as a signature pattern for these enzymes. As a second signature pattern a conserved region was selected in the C-terminal part of the PEP-utilizing enzymes. The biological significance of this region is not yet known.

Consensus pattern: G-[GA]-x-[TN]-x-H-[STA]-[STAV SEQ ID NO:105)]-[LIVM SEQ ID NO:4)](2)-[STAV SEQ ID NO:105)]-[RG] [H is phosphorylated]
-Consensus pattern: [DEQSK SEQ ID NO:402)]-x-[LIVMF SEQ ID NO:2)]-S-[LIVMF SEQ ID NO:2)]-G-[ST]-N-D-[LIVM SEQ ID NO:4)]-x-Q-[LIVMFYGT SEQ ID NO:403)]-[STALIV SEQ ID NO:404)]-[LIVMF SEQ ID NO:2)]-[GAS]-x(2)-R

[1] Reizer J., Hoischen C., Reizer A., Pham T.N., Saier M.H. Jr. Protein Sci. 2:506-521(1993).

[2] Reizer J., Reizer A., Merrick M.J., Plunkett G. III, Rose D.J., Saier M.H. Jr. Gene 181:103-108(1996).

[3] Pocalyko D.J., Carroll L.J., Martin B.M., Babbitt P.C., Dunaway-Mariano D. Biochemistry 29:10757-10765(1990).

[4] Niersbach M., Kreuzaler F., Geerse R.H., Postma P., Hirsch H.J. Mol. Gen. Genet. 232:332-336(1992).

402. (PEPCK ATP) Phosphoenolpyruvate carboxykinase (ATP) signature

Phosphoenolpyruvate carboxykinase (ATP) (EC 4.1.1.49) (PEPCK) [1] catalyzes the formation of phosphoenolpyruvate by decarboxylation of oxaloacetate while hydrolyzing ATP, a rate limiting step in gluconeogenesis (the biosynthesis of glucose).

The sequence of this enzyme has been obtained from *Escherichia coli*, yeast, and *Trypanosoma brucei*; these three sequences are evolutionary related and share many regions of similarity. As a signature pattern a highly conserved region was selected that contains four acidic residues and which is located in the central part of the enzyme. The beginning of the pattern is located about 10 residues to the C-terminus of an ATP-binding motif 'A' (P-loop) (see <PDOC00017>) and is also part of the ATP-binding domain [2].

Consensus pattern: L-I-G-D-D-E-H-x-W-x-[DE]-x-G-[IV]-x-N

-Note: phosphoenolpyruvate carboxykinase (GTP) (EC 4.1.1.32) an enzyme that catalyzes the same reaction, but using GTP instead of ATP, is not related to the above enzyme (see <PDOC00421>).

[1] Medina V., Pontarollo R., Glaeske D., Tabel H., Goldie H. J. Bacteriol. 172:7151-7156(1990).

[2] Matte A., Goldie H., Sweet R.M., Delbaere L.T.J. J. Mol. Biol. 256:126-143(1996).

403. (Pepcase) Phosphoenolpyruvate carboxylase active sites. Phosphoenolpyruvate carboxylase (EC 4.1.1.31) (PEPcase) catalyzes the irreversible beta-carboxylation of phosphoenolpyruvate by bicarbonate to yield oxaloacetate and phosphate. The enzyme is found in all plants and in a variety of microorganisms. A histidine [1] and a lysine [2] have been implicated in the catalytic mechanism of this enzyme; the regions around these active site residues are highly conserved in PEPcase from various plants, bacteria and cyanobacteria and can be used as a signature patterns for this type of enzyme.

Consensus pattern: [VT]-x-T-A-H-P-T-[EQ]-x(2)-R-[KRH] [H is an active site residue]-

Consensus pattern: [IV]-M-[LIVM SEQ ID NO:4)]-G-Y-S-D-S-x-K-D-[STAG SEQ ID NO:20)]-G [K is an active site residue]-

- 5 [1] Terada K., Izui K. Eur. J. Biochem. 202:797-803(1991).[2] Jiao J.-A., Podesta F.E., Chollet R., O'Leary M.H., Andreo C.S. Biochim. Biophys. Acta 1041:291-295(1990).

404. PET112 family signature

10 The following proteins from eukaryotes, prokaryotes and archaeobacteria belong to the same family:

- Yeast mitochondrial protein PET112 [1], which plays an unknown role in the expression of mitochondrial genes, probably at the level of translation.
- Aspergillus nidulans mitochondrial protein nempA.
- 15 - Bacillus subtilis hypothetical protein yzdD.
- Moraxella catarrhalis hypothetical protein in bloR-1 3'region.
- Mycoplasma genitalium hypothetical protein MG100.
- Methanococcus jannaschii hypothetical proteins MJ0019 and MJ0160.

20 The size of these proteins range from 419 to 630 amino acids. As a signature pattern, a conserved region located in the N-terminal section was selected.

Consensus pattern: [DN]-x-[DN]-R-x(3)-P-L-[LIV]-E-[LIV]-x-[ST]-x-P

- 25 [1] Mulero J.J., Rosenthal J.K., Fox T.D. Curr. Genet. 25:299-304(1994).

405. (PFK) Phosphofructokinase signature

Phosphofructokinase (EC 2.7.1.11) (PFK) [1,2] is a key regulatory enzyme in the glycolytic pathway. It catalyzes the phosphorylation by ATP of fructose 30 6-phosphate to fructose 1,6-bisphosphate. In bacteria PFK is a tetramer of identical 36 Kd subunits. In mammals it is a tetramer of 80 Kd subunits. Each 80 Kd subunit consist of two homologous domains which are highly related to the bacterial 36 Kd subunits. In Human there are three, tissue-specific, types

of PFK isozymes: PFKM (muscle), PFKL (liver), and PFKP (platelet). In yeast PFK is an octamer composed of four 100 Kd alpha chains (gene PFK1) and four 100 Kd beta chains (gene PFK2); like the mammalian 80 Kd subunits, the yeast 100 Kd subunits are composed of two homologous domains.

- 5 As a signature pattern for PFK a region that contains three basic residues involved in fructose-6-phosphate binding was selected.

Consensus pattern: [RK]-x(4)-G-H-x-Q-[QR]-G-G-x(5)-D-R [The R/K, the H and the Q/R are involved in fructose-6-P binding]

- 10 -Note: Escherichia coli has two phosphofructokinase isozymes which are encoded by genes pfkA (major) and pfkB (minor). The pfkB isozyme is not evolutionary related to other prokaryotic or eukaryotic PFK's (see <PDOC00504>).

[1] Poorman R.A., Randolph A., Kemp R.G., Henrikson R.L. Nature 309:467-469(1984).

- 15 [2] Heinisch J., Ritzel R.G., von Borstel R.C., Aguilera A., Rodicio R., Zimmermann F.K. Gene 78:309-321(1989).

406. (PGAM) Phosphoglycerate mutase family phosphohistidine signature

- 20 Phosphoglycerate mutase (EC 5.4.2.1) (PGAM) and bisphosphoglycerate mutase (EC 5.4.2.4) (BPGM) are structurally related enzymes which catalyze reactions involving the transfer of phospho groups between the three carbon atoms of phosphoglycerate [1,2]. Both enzymes can catalyze three different reactions, although in different proportions:

- 25 - The isomerization of 2-phosphoglycerate (2-PGA) to 3-phosphoglycerate (3-PGA) with 2,3-diphosphoglycerate (2,3-DPG) as the primer of the reaction.
- The synthesis of 2,3-DPG from 1,3-DPG with 3-PGA as a primer.
- The degradation of 2,3-DPG to 3-PGA (phosphatase EC 3.1.3.13 activity).

- In mammals, PGAM is a dimeric protein. There are two isoforms of PGAM: the M (muscle) and B (brain) forms. In yeast, PGAM is a tetrameric protein. BPGM is a dimeric protein and is found mainly in erythrocytes where it plays a major role in regulating hemoglobin oxygen affinity as a consequence of controlling 2,3-DPG concentration.
- 30

The catalytic mechanism of both PGAM and BPGM involves the formation of a phosphohistidine intermediate [3].

The bifunctional enzyme 6-phosphofructo-2-kinase / fructose-2,6-bisphosphatase (EC 2.7.1.105 and EC 3.1.3.46) (PF2K) [4] catalyzes both the synthesis and the

5 degradation of fructose-2,6-bisphosphate. PF2K is an important enzyme in the regulation of hepatic carbohydrate metabolism. Like PGAM/BPGM, the fructose-2,6-bisphosphatase reaction involves a phosphohistidine intermediate and the phosphatase domain of PF2K is structurally related to PGAM/BPGM.

The bacterial enzyme alpha-ribazole-5'-phosphate phosphatase (gene cobC) which
10 is involved in cobalamin biosynthesis also belongs to this family [5].

A signature pattern was built around the phosphohistidine residue.

Consensus pattern: [LIVM SEQ ID NO:4)]-x-R-H-G-[EQ]-x(3)-N [H is the phosphohistidine residue]

15 -Note: some organisms harbor a form of PGAM independent of 2,3-DPG, this enzyme is not related to the family described above [6].

[1] Le Boulch P., Joulin V., Garel M.-C., Rosa J., Cohen-Solal M. Biochem. Biophys. Res. Commun. 156:874-881(1988).

20 [2] White M.F., Fothergill-Gilmore L.A. FEBS Lett. 229:383-387(1988).

[3] Rose Z.B. Meth. Enzymol. 87:43-51(1982).

[4] Bazan J.F., Fletterick R.J., Pilkis S.J. Proc. Natl. Acad. Sci. U.S.A. 86:9642-9646(1989).

[5] O'Toole G.A., Trzebiatowski J.R., Escalante-Semerena J.C. J. Biol. Chem. 269:26503-26511(1994).

25 [6] Grana X., De Lecea L., El-Maghrabi M.R., Urena J.M., Caellas C., Carreras J., Puigdomenech P., Pilkis S.J., Climent F. J. Biol. Chem. 267:12797-12803(1992).

30 407. (PGI) Phosphoglucose isomerase signatures

Phosphoglucose isomerase (EC 5.3.1.9) (PGI) [1,2] is a dimeric enzyme that catalyzes the reversible isomerization of glucose-6-phosphate and fructose-6-phosphate. PGI is involved in different pathways: in most higher organisms it

is involved in glycolysis; in mammals it is involved in gluconeogenesis; in plants in carbohydrate biosynthesis; in some bacteria it provides a gateway for fructose into the Entner-Doudouroff pathway. PGI has been shown [3] to be identical to neuroleukin, a neurotrophic factor which supports the survival of various types of neurons.

The sequence of PGI from many species ranging from bacteria to mammals is available and has been shown to be highly conserved. As signature patterns for this enzyme two conserved regions were selected, the first region is located in the central section of PGI, while the second one is located in its C-terminal section.

Consensus pattern: [DENS SEQ ID NO:405)]-x-[LIVM SEQ ID NO:4)]-G-G-R-[FY]-S-[LIVMT SEQ ID NO:1)]-x-[STA]-[PSAC SEQ ID NO:406)]-[LIVMA SEQ ID NO:30)]-G-Consensus pattern: [GS]-x-[LIVM SEQ ID NO:4)]-[LIVMFYW SEQ ID NO:26)]-x(4)-[FY]-[DN]-Q-x-G-V-E-x(2)-K

[1] Achari A., Marshall S.E., Muirhewad H., Palmieri R.H., Noltmann E.A. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 293:145-157(1981).

[2] Smith M.W., Doolittle R.F. J. Mol. Evol. 34:544-545(1992).

[3] Faik P., Walker J.I.H., Redmill A.A.M., Morgan M.J. Nature 332:455-456(1988).

408. (PGK) Phosphoglycerate kinase signature

Phosphoglycerate kinase (EC 2.7.2.3) (PGK) [1] catalyzes the second step in the second phase of glycolysis, the reversible conversion of 1,3-diphosphoglycerate to 3-phosphoglycerate with generation of one molecule of ATP. PGK is found in all living organisms and its sequence has been highly conserved throughout evolution. It is a two-domain protein; each domain is composed of six repeats of an alpha/beta structural motif. As a signature pattern for PGK's, a conserved region in the N-terminal region was selected.

Consensus pattern: [KRHGTCVN SEQ ID NO:407)]-[VT]-[LIVMF SEQ ID NO:2)]-[LIVMC SEQ ID NO:142)]-R-x-D-x-N-[SACV SEQ ID NO:391)]-P

[1] Watson H.C., Littlechild J.A. Biochem. Soc. Trans. 18:187-190(1990).

409. (PGM PMM) Phosphoglucomutase and phosphomannomutase phosphoserine signature

- 5 - Phosphoglucomutase (EC 5.4.2.2) (PGM). PGM is an enzyme responsible for the conversion of D-glucose 1-phosphate into D-glucose 6-phosphate. PGM participates in both the breakdown and synthesis of glucose [1].
- Phosphomannomutase (EC 5.4.2.8) (PMM). PMM is an enzyme responsible for the conversion of D-mannose 1-phosphate into D-mannose 6-phosphate. PMM is
10 required for different biosynthetic pathways in bacteria. For example, in enterobacteria such as *Escherichia coli* there are two different genes coding for this enzyme: *rfbK* which is involved in the synthesis of the O antigen of lipopolysaccharide and *cpsG* which is required for the synthesis of the M antigen capsular polysaccharide [2]. In *Pseudomonas aeruginosa* PMM
15 (gene *algC*) is involved in the biosynthesis of the alginate layer [3] and in *Xanthomonas campestris* (gene *xanA*) it is involved in the biosynthesis of xanthan [4]. In *Rhizobium* strain ngr234 (gene *noeK*) it is involved in the biosynthesis of the nod factor.
- Phosphoacetylglucosamine mutase (EC 5.4.2.3) which converts N-acetyl-D-glucosamine 1-phosphate into the 6-phosphate isomer.
20

The catalytic mechanism of both PGM and PMM involves the formation of a phosphoserine intermediate [1]. The sequence around the serine residue is well conserved and can be used as a signature pattern.

In addition to PGM and PMM there are at least three uncharacterized proteins
25 that belong to this family [5,6]:

- Urease operon protein *ureC* from *Helicobacter pylori*.
- *Escherichia coli* protein *mrsA*.
- *Paramecium tetraurelia* *parafusin*, a phosphoglycoprotein involved in exocytosis.
- 30 - A *Methanococcus vannielii* hypothetical protein in the 3' region of the gene for ribosomal protein S10.

Consensus pattern: [GSA]-[LIVM SEQ ID NO:4)]-x-[LIVM SEQ ID NO:4)]-[ST]-[PGA]-S-H-x-P-x(4)-[GNHE SEQ ID NO:408)] [S is the phosphoserine residue]

-Note: PMM from fungi do not belong to this family.

[1] Dai J.B., Liu Y., Ray W.J. Jr., Konno M. J. Biol. Chem. 267:6322-6337(1992).

[2] Stevenson G., Lee S.J., Romana L.K., Reeves P.R. Mol. Gen. Genet. 227:173-180(1991).

[3] Zielinski N.A., Chakrabarty A.M., Berry A. J. Biol. Chem. 266:9754-9763(1991).

[4] Koeplin R., Arnold W., Hoette B., Simon R., Wang G., Puehler A. J. Bacteriol.

174:191-199(1992).

[5] Bairoch A. Unpublished observations (1993).

[6] Subramanian S.V., Wyroba E., Andersen A.P., Satir B.H. Proc. Natl. Acad. Sci. U.S.A. 91:9832-9836(1994).

410. PH domain profile

The 'pleckstrin homology' (PH) domain is a domain of about 100 residues that occurs in a wide range of proteins involved in intracellular signaling or as constituents of the cytoskeleton [1 to 7].

The function of this domain is not clear, several putative functions have been suggested: - binding to the beta/gamma subunit of heterotrimeric G proteins,

- binding to lipids, e.g. phosphatidylinositol-4,5-bisphosphate,

- binding to phosphorylated Ser/Thr residues,

- attachment to membranes by an unknown mechanism.

It is possible that different PH domains have totally different ligand requirements.

The 3D structure of several PH domains has been determined [8]. All known cases have a common structure consisting of two perpendicular anti-parallel beta sheets, followed by a C-terminal amphipathic helix. The loops connecting the beta-strands differ greatly in length, making the PH domain relatively difficult to detect. There are no totally invariant residues within the PH domain.

Proteins reported to contain one more PH domains belong to the following

families:

- Pleckstrin, the protein where this domain was first detected, is the major substrate of protein kinase C in platelets. Pleckstrin is one of the rare proteins to contains two PH domains.
- 5 - Ser/Thr protein kinases such as the Act/Rac family, the beta-adrenergic receptor kinases, the mu isoform of PKC and the trypanosomal NrK1 family.
- Tyrosine protein kinases belonging to the Btk/Itk/Tec subfamily.
- Insulin Receptor Substrate 1 (IRS-1).
- Regulators of small G-proteins like guanine nucleotide releasing factor
- 10 GNRP (Ras-GRF) (which contains 2 PH domains), guanine nucleotide exchange proteins like vav, dbp, Sos and yeast CDC24, GTPase activating proteins like rasGAP and BEM2/IPL2, and the human breakpoint cluster protein bcr.
- Cytoskeletal proteins such as dynamin (see <PDOC00362>), Caenorhabditis elegans kinesin-like protein unc-104 (see <PDOC00343>), spectrin beta-chain, syntrophin (2 PH domains) and yeast nuclear migration protein NUM1.
- 15 - Mammalian phosphatidylinositol-specific phospholipase C (PI-PLC) (see <PDOC50007>) isoforms gamma and delta. Isoform gamma contains two PH domains, the second one is split into two parts separated by about 400 residues.
- Oxysterol binding proteins OSBP, yeast OSH1 and YHR073w.
- 20 - Mouse protein citron, a putative rho/rac effector that binds to the GTP-bound forms of rho and rac,
- Several yeast proteins involved in cell cycle regulation and bud formation like BEM2, BEM3, BUD4 and the BEM1-binding proteins BOI2 (BEB1) and BOI1 (BOB1).
- Caenorhabditis elegans protein MIG-10.
- 25 - Caenorhabditis elegans hypothetical proteins C04D8.1, K06H7.4 and ZK632.12.
- Yeast hypothetical proteins YBR129c and YHR155w.

The profile for the PH domain, which has been developed by Toby Gibson at the EMBL, covers the total length of domain. Several proteins contain large insertions in the PH domain and are thus difficult to detect with this

30 profile. In some of these cases, the profile will align only to one half of the PH domain.

-Sequences known to belong to this class detected by the pattern: ALL. But it should be noted that while all sequences containing PH domains are detected,

not all PH domains are. Some of the split domains lie below the cutoff threshold.

[1] Mayer B.J., Ren R., Clark K.L., Baltimore D. Cell 73:629-630(1993).

[2] Haslam R.J., Koide H.B., Hemmings B.A. Nature 363:309-310(1993).

5 [3] Musacchio A., Gibson T.J., Rice P., Thompson J., Saraste M.

Trends Biochem. Sci. 18:343-348(1993).

[4] Gibson T.J., Hyvonen M., Musacchio A., Saraste M., Birney E.

Trends Biochem. Sci. 19:349-353(1994).[5] Pawson T.

Nature 373:573-580(1995).[6] Ingley E., Hemmings B.A.

10 J. Cell. Biochem. 56:436-443(1994).[7] Saraste M., Hyvonen M.

Curr. Opin. Struct. Biol. 5:403-408(1995).[8] Riddihough G.

Nat. Struct. Biol. 1:755-757(1994).

15 411. PHD-finger

[1]

Medline: 95216093

The PHD finger: implications for chromatin-mediated transcriptional regulation.

20 Aasland R, Gibson TJ, Stewart AF;

Trends Biochem Sci 1995;20:56-59.

Number of members: 181

25 412. (PI-PLC-X) Phosphatidylinositol-specific phospholipase C profiles

Phosphatidylinositol-specific phospholipase C (EC 3.1.4.11), an eukaryotic intracellular enzyme, plays an important role in signal transduction processes

[1]. It catalyzes the hydrolysis of 1-phosphatidyl-D-myo-inositol-3,4,5-triphosphate into the second messenger molecules diacylglycerol and inositol-

30 1,4,5-triphosphate. This catalytic process is tightly regulated by reversible phosphorylation and binding of regulatory proteins [2 to 4].

In mammals, there are at least 6 different isoforms of PI-PLC, they differ in their domain structure, their regulation, and their tissue distribution. Lower

eukaryotes also possess multiple isoforms of PI-PLC.

All eukaryotic PI-PLCs contain two regions of homology, sometimes referred to as 'X-box' and 'Y-box'. The order of these two regions is always the same (NH₂-X-Y-COOH), but the spacing is variable. In most isoforms, the distance between these two regions is only 50-100 residues but in the gamma isoforms one PH domain, two SH2 domains, and one SH3 domain are inserted between the two PLC-specific domains. The two conserved regions have been shown to be important for the catalytic activity. At the C-terminal of the Y-box, there is a C2 domain (see <PDOC00380>) possibly involved in Ca-dependent membrane attachment.

Profile analysis shows that sequences with significant similarity to the X-box domain occur also in prokaryotic and trypanosome PI-specific phospholipases C. Apart from this region, the prokaryotic enzymes show no similarity to their eukaryotic counterparts.

Two profiles were developed, one covering the X-box, the other the Y-box.

[1] Meldrum E., Parker P.J., Carozzi A.

Biochim. Biophys. Acta 1092:49-71(1991).[2] Rhee S.G., Choi K.D.

Adv. Second Messenger Phosphoprotein Res. 26:35-61(1992).

[3] Rhee S.G., Choi K.D. J. Biol. Chem. 267:12393-12396(1992).

[4] Sternweis P.C., Smrcka A.V. Trends Biochem. Sci. 17:502-506(1992).

413. (PI-PLC-Y) Phosphatidylinositol-specific phospholipase C profiles

Phosphatidylinositol-specific phospholipase C (EC 3.1.4.11), an eukaryotic

intracellular enzyme, plays an important role in signal transduction processes

[1]. It catalyzes the hydrolysis of 1-phosphatidyl-D-myo-inositol-3,4,5-triphosphate into the second messenger molecules diacylglycerol and inositol-1,4,5-triphosphate. This catalytic process is tightly regulated by reversible phosphorylation and binding of regulatory proteins [2 to 4].

In mammals, there are at least 6 different isoforms of PI-PLC, they differ in their domain structure, their regulation, and their tissue distribution. Lower eukaryotes also possess multiple isoforms of PI-PLC.

All eukaryotic PI-PLCs contain two regions of homology, sometimes referred to

as 'X-box' and 'Y-box'. The order of these two regions is always the same (NH₂-X-Y-COOH), but the spacing is variable. In most isoforms, the distance between these two regions is only 50-100 residues but in the gamma isoforms one PH domain, two SH2 domains, and one SH3 domain are inserted between the two PLC-specific domains. The two conserved regions have been shown to be important for the catalytic activity. At the C-terminal of the Y-box, there is a C2 domain (see <PDOC00380>) possibly involved in Ca-dependent membrane attachment.

Profile analysis shows that sequences with significant similarity to the X-box domain occur also in prokaryotic and trypanosome PI-specific phospholipases C. Apart from this region, the prokaryotic enzymes show no similarity to their eukaryotic counterparts.

Two profiles were developed, one covering the X-box, the other the Y-box.

[1] Meldrum E., Parker P.J., Carozzi A.

Biochim. Biophys. Acta 1092:49-71(1991).[2] Rhee S.G., Choi K.D.

Adv. Second Messenger Phosphoprotein Res. 26:35-61(1992).

[3] Rhee S.G., Choi K.D. J. Biol. Chem. 267:12393-12396(1992).

[4] Sternweis P.C., Smrcka A.V. Trends Biochem. Sci. 17:502-506(1992).

414. (PK) Pyruvate kinase active site signature

Pyruvate kinase (EC 2.7.1.40) (PK) [1] catalyzes the final step in glycolysis, the conversion of phosphoenolpyruvate to pyruvate with the concomitant phosphorylation of ADP to ATP. PK requires both magnesium and potassium ions for its activity. PK is found in all living organisms. In vertebrates there are four, tissues specific, isozymes: L (liver), R (red cells), M1 (muscle, heart, and brain), and M2 (early fetal tissues). In Escherichia coli there are two isozymes: PK-I (gene pykF) and PK-II (gene pykA). All PK isozymes seem to be tetramers of identical subunits of about 500 amino acid residues.

As a signature pattern for PK a conserved region was selected that includes a lysine residue which seems to be the acid/base catalyst responsible for the interconversion of pyruvate and enolpyruvate, and a glutamic acid residue implicated in the binding of the magnesium ion.

Consensus pattern: [LIVAC SEQ ID NO:319])-x-[LIVM SEQ ID NO:4)](2)-[SAPCV SEQ ID NO:409)]-K-[LIV]-E-[NKRST SEQ ID NO:410)]-x-[DEQHS SEQ ID NO:411)]-[GSTA SEQ ID NO:19)]-[LIVM SEQ ID NO:4)] [K is the active site residue] [E is a magnesium
5 ligand]

[1] Muirhead H. Biochem. Soc. Trans. 18:193-196(1990).

10 415. (PLDc) Phospholipase D. Active site motif

Phosphatidylcholine-hydrolyzing phospholipase D (PLD) isoforms are activated by ADP-ribosylation factors (ARFs). PLD produces phosphatidic acid from phosphatidylcholine, which may be essential for the formation of certain types of transport vesicles or may be constitutive vesicular
15 transport to signal transduction pathways.

PC-hydrolyzing PLD is a homologue of cardiolipin synthase, phosphatidylserine synthase, bacterial PLDs, and viral proteins.

Each of these appears to possess a domain duplication which is apparent by the presence of two motifs containing well-conserved histidine, lysine,
20 and/or asparagine residues which may contribute to the active site. aspartic acid. An E. coli endonuclease (nuc) and similar proteins appear to be PLD homologues but possess only one of these motifs.

The profile contained here represents only the putative active site regions, since an accurate multiple alignment of the repeat units
25 has not been achieved.

Number of members: 139

[1]

Medline: 96303814

A novel family of phospholipase D homologues that includes
30 phospholipid synthases and putative endonucleases:
identification of duplicated repeats and potential active site residues.

Ponting CP, Kerr ID;

Protein Sci 1996;5:914-922.

[2]Medline: 96334293

A duplicated catalytic motif in a new superfamily of phosphohydrolases and phospholipid synthases that includes poxvirus envelope proteins.

Koonin EV;

Trends Biochem Sci 1996;21:242-243.

[3]Medline: 94327597

Cloning and expression of phosphatidylcholine-hydrolyzing phospholipase D from *Ricinus communis* L.

Wang X, Xu L, Zheng L;

J Biol Chem 1994;269:20312-20317.

[4]Medline: 97386825

Regulation of eukaryotic phosphatidylinositol-specific phospholipase C and phospholipase D.

Singer WD, Brown HA, Sternweis PC;

Annu Rev Biochem 1997;66:475-509.

416. (PMI type1) Phosphomannose isomerase type I signatures

Phosphomannose isomerase (EC 5.3.1.8) (PMI) [1,2] is the enzyme that catalyzes the interconversion of mannose-6-phosphate and fructose-6-phosphate. In eukaryotes, it is involved in the synthesis of GDP-mannose which is a constituent of N- and O-linked glycans as well as GPI anchors. In prokaryotes, it is involved in a variety of pathways including capsular polysaccharide biosynthesis and D-mannose metabolism.

Three classes of PMI have been defined on the basis of sequence similarities [1]. The first class comprises all known eukaryotic PMI as well as the enzyme encoded by the *manA* gene in enterobacteria such as *Escherichia coli*. Class I PMI's are proteins of about 42 to 50 Kd which bind a zinc ion essential for their activity.

As signature patterns for class I PMI, two conserved regions were selected. The first one is located in the N-terminal section of these proteins, the second

in the C-terminal half. Both patterns contain a residue involved [3] in the binding of the zinc ion.

Consensus pattern: Y-x-D-x-N-H-K-P-E [E is a zinc ligand]

5 -Consensus pattern: H-A-Y-[LIVM SEQ ID NO:4)]-x-G-x(2)-[LIVM SEQ ID NO:4)]-E-x-M-A-x-S-D-N-x-[LIVM SEQ ID NO:4)]-R-A-G-x-T-P-K [H is a zinc ligand]

[1] Proudfoot A.E.I., Turcatti G., Wells T.N.C., Payton M.A., Smith D.J. Eur. J. Biochem. 219:415-423(1994).

10 [2] Coulin F., Magnenat E., Proudfoot A.E.I., Payton M.A., Scully P., Wells T.N.C. Biochemistry 32:14139-14144(1993).

[3] Cleasby A., Wonacott A., Skarzynski T., Hubbard R.E., Davies G.J., Proudfoot A.E.I., Bernard A.R., Payton M.A., Wells T.N.C. Nat. Struct. Biol. 3:470-479(1996).

15

417. (PNP UDP 1) Purine and other phosphorylases family 1 signature

The following phosphorylases belongs to the same family:

- Purine nucleoside phosphorylase (EC 2.4.2.1) (PNP) from most bacteria (gene deoD). This enzyme catalyzes the cleavage of guanosine or inosine to
20 respective bases and sugar-1-phosphate molecules [1].
- Uridine phosphorylase (EC 2.4.2.3) (UdRPase) from bacteria (gene udp) and mammals. Catalyzes the cleavage of uridine into uracil and ribose-1-phosphate. The products of the reaction are used either as carbon and energy sources or in the rescue of pyrimidine bases for nucleotide
25 synthesis [2].
- 5'-methylthioadenosine phosphorylase (EC 2.4.2.28) (MTA phosphorylase) from *Sulfolobus solfataricus* [3].

As a signature pattern, a conserved region was selected in the central part of these enzymes.

30

Consensus pattern: [GST]-x-G-[LIVM SEQ ID NO:4)]-G-x-[PA]-S-x-[GSTA SEQ ID NO:19)]-I-x(3)-E-L

-Note: it should be noted that mammalian and some bacterial PNP as well as eukaryotic MTA phosphorylase belong to a different family of phosphorylases (see <PDOC00954>).

[1] Takehara M., Ling F., Izawa S., Inoue Y., Kimura A. Biosci. Biotechnol. Biochem.
5 59:1987-1990(1995).

[2] Watanabe S.-I., Hino A., Wada K., Eliason J.F., Uchida T. J. Biol. Chem. 270:12191-
12196(1995).

[3] Cacciapuoti G., Porcelli M., Bertoldo C., De Rosa M., Zappia V. J. Biol. Chem.
269:24762-24769(1994).

10

418. (PP2C) Protein phosphatase 2C signature

Protein phosphatase 2C (PP2C) is one of the four major classes of mammalian
serine/threonine specific protein phosphatases (EC 3.1.3.16). PP2C [1] is a
15 monomeric enzyme of about 42 Kd which shows broad substrate specificity and
is dependent on divalent cations (mainly manganese and magnesium) for its
activity. Its exact physiological role is still unclear. Three isozymes are
currently known in mammals: PP2C-alpha, -beta and -gamma. In yeast, there are
at least four PP2C homologs: phosphatase PTC1 [2] which has weak tyrosine
20 phosphatase activity in addition to its activity on serines, phosphatases PTC2
and PTC3, and hypothetical protein YBR125c. Isozymes of PP2C are also known
from *Arabidopsis thaliana* (ABI1, PPH1), *Caenorhabditis elegans* (FEM-2,
F42G9.1, T23F11.1), *Leishmania chagasi* and *Paramecium tetraurelia*.

In *Arabidopsis thaliana*, the kinase associated protein phosphatase (KAPP) [3]
25 is an enzyme that dephosphorylates the Ser/Thr receptor-like kinase RLK5 and
which contains a C-terminal PP2C domain.

PP2C does not seem to be evolutionary related to the main family of serine/
threonine phosphatases: PP1, PP2A and PP2B. However, it is significantly
similar to the catalytic subunit of pyruvate dehydrogenase phosphatase
30 (EC 3.1.3.43) (PDPC) [4], which catalyzes dephosphorylation and concomitant
reactivation of the alpha subunit of the E1 component of the pyruvate
dehydrogenase complex. PDPC is a mitochondrial enzyme and, like PP2C, is
magnesium-dependent.

As a signature pattern, the best conserved region was selected which is located in the N-terminal part and contains a perfectly conserved tripeptide. This region includes a conserved aspartate residue involved in divalent cation binding [5].

5

Consensus pattern: [LIVMFY SEQ ID NO:18)]-[LIVMFYA SEQ ID NO:98)]-[GSAC SEQ ID NO:93)]-[LIVM SEQ ID NO:4)]-[FYC]-D-G-H-[GAV]

-Note: PP2C belongs [6] to a superfamily which also includes bacterial proteins such as *Bacillus spoIIE*, *rsbU* and *rsbW*, *Synechocystis* PCC 6803 *icfG* as well as a domain in fungal adenylate cyclases.

10

[1] Wenk J., Trompeter H.-I., Pettrich K.-G., Cohen P.T.W., Campbell D.G., Mieskes G. *FEBS Lett.* 297:135-138(1992).

[2] Maeda T., Tsai A.Y.M., Saito H. *Mol. Cell. Biol.* 13:5408-5417(1993).

15

[3] Stone J.M., Collinge M.A., Smith R.D., Horn M.A., Walker J.C. *Science* 266:793-795(1994).

[4] Lawson J.E., Niu X.-D., Browning K.S., Trong H.L., Yan J., Reed L.J. *Biochemistry* 32:8987-8993(1993).

[5] Das A.K., Helps N.R., Cohen P.T.W., Barford D. *EMBO J.* 24:6798-6809(1996).

20

[6] Bork P., Brown N.P., Hegyi H., Schultz J. *Protein Sci.* 5:1421-1425(1996).

419. (PPTA) Protein prenyltransferases alpha subunit repeat signature

Protein prenyltransferases catalyze the transfer of an isoprenyl moiety to a cysteine four residues from the C-terminus of several proteins. They are heterodimeric enzymes consisting of alpha and beta subunits. The alpha subunit is thought to participate in a stable complex with the isoprenyl substrate; the beta subunit binds the peptide substrate. Distinct protein prenyltransferases might share a common alpha subunit. Both the alpha and beta subunit show repetitive sequence motifs [1]. These repeats have distinct structural and functional implications and are unrelated to each other. Known protein prenyltransferase alpha subunits are:

25

30

- Mammalian protein farnesyltransferase alpha subunit.

- Yeast protein RAM2, a protein farnesyltransferase alpha subunit.
- Yeast protein BET4, a protein geranylgeranyltransferase alpha subunit.

The conserved domain of the alpha subunit consists of about 34 amino acids and is repeated five times. It contains an invariant tryptophan possibly involved in heterodimerization with the conserved phenylalanines in the repeated domains of the beta subunits, via hydrophobic bonds. The signature pattern for this domain is centered on the invariant tryptophan.

Consensus pattern: [PSIAV SEQ ID NO:412)]-x-[NDFV SEQ ID NO:413)]-[NEQIY SEQ ID NO:414)]-x-[LIVMAGP SEQ ID NO:415)]-W-[NQSTHF SEQ ID NO:416)]-[FYHQ SEQ ID NO:417)]-[LIVMR SEQ ID NO:418)]

[1] Boguski M.S., Murray A.W., Powers S. New Biol. 4:408-411(1992).

420. (PR55) Protein phosphatase 2A regulatory subunit PR55 signatures

Protein phosphatase 2A (PP2A) is a serine/threonine phosphatase involved in many aspects of cellular function including the regulation of metabolic enzymes and proteins involved in signal transduction. PP2A is a trimeric enzyme that consists of a core composed of a catalytic subunit associated with a 65 Kd regulatory subunit (PR65), also called subunit A; this complex then associates with a third variable subunit (subunit B), which confers distinct properties to the holoenzyme [1]. One of the forms of the variable subunit is a 55 Kd protein (PR55) which is highly conserved in mammals - where three isoforms are known to exist -, Drosophila and yeast (gene CDC55). This subunit could perform a substrate recognition function or be responsible for targeting the enzyme complex to the appropriate subcellular compartment.

As signature patterns, two perfectly conserved sequences of 15 residues were selected; one located in the N-terminal region, the other in the center of the protein.

Consensus pattern: E-F-D-Y-L-K-S-L-E-I-E-E-K-I-N

Consensus pattern: N-[AG]-H-[TA]-Y-H-I-N-S-I-S-[LIVM SEQ ID NO:4)]-N-S-D

[1] Mayer-Jackel R., Hemmings B.A. Trends Cell Biol. 4:287-291(1994).

5 421. N-(5'phosphoribosyl)anthranilate (PRA) isomerase

[1] Wilmanns M, Priestle JP, Niermann T, Jansonius JN;
J Mol Biol 1992;223:477-507.

10 422. (PRK) Phosphoribulokinase signature

Phosphoribulokinase (EC 2.7.1.19) (PRK) [1,2] is one of the enzymes specific to the Calvin's reductive pentose phosphate cycle which is the major route by which carbon dioxide is assimilated and reduced by autotrophic organisms. PRK catalyzes the ATP-dependent phosphorylation of ribulose 5-phosphate into ribulose 1,5-bisphosphate which is the substrate for RubisCO.

15 PRK's of diverse origins show different properties with respect to the size of the protein, the subunit structure, or the enzymatic regulation. However an alignment of the sequences of PRK from plants, algae, photosynthetic and chemoautotrophic bacteria shows that there are a few regions of sequence
20 similarity. As a signature pattern one of these regions was selected.

Consensus pattern: K-[LIVM SEQ ID NO:4)]-x-R-D-x(3)-R-G-x-[ST]-x-E

[1] Kossmann J., Klintworth R., Bowien B. Gene 85:247-252(1989).

25 [2] Gibson J.L., Chen J.-H., Tower P.A., Tabita F.R. Biochemistry 29:8085-8093(1990).

423. (PRPP synt) Phosphoribosyl pyrophosphate synthetase signature

Phosphoribosyl pyrophosphate synthetase (EC 2.7.6.1) (PRPP synthetase)
30 catalyzes the formation of PRPP from ATP and ribose 5-phosphate. PRPP is then used in various biosynthetic pathways, as for example in the formation of purines, pyrimidines, histidine and tryptophan. PRPP synthetase requires inorganic phosphate and magnesium ions for its stability and activity.

In mammals, three isozymes of PRPP synthetase are found; in yeast there are at least four isozymes.

As a signature pattern for this enzyme, a very conserved region was selected that has been suggested to be involved in binding divalent cations [1]. This region contains two conserved aspartic acid residues as well as a histidine, which are all potential ligands for a cation such as magnesium.

Consensus pattern: D-[LI]-H-[SA]-x-Q-[IMST SEQ ID NO:419)]-[QM]-G-[FY]-F-x(2)-P-[LIVMFC SEQ ID NO:90)]-D

[1] Bower S.G., Harlow K.W., Switzer R.L., Hoven-Jensen B. J. Biol. Chem. 264:10287-10291(1989).

424. (PRTP) Herpesvirus processing and transport protein

The members of this family are associated with capsid intermediates during packaging of the virus.

Number of members: 31

[1]

Medline: 98362148

Herpes simplex virus type 1 cleavage and packaging proteins
UL15 and UL28 are associated with B but not C capsids during packaging. Yu D, Weller SK;
J Virol 1998;72:7428-7439.

425. Photosystem I psaG / psaK (PSI PSAK) proteins signature

Photosystem I (PSI) [1] is an integral membrane protein complex that uses light energy to mediate electron transfer from plastocyanin to ferredoxin. It is found in the chloroplasts of plants and cyanobacteria. PSI is composed of at least 14 different subunits, two of which PSI-G (gene psaG) and PSI-K (gene psaK) are small hydrophobic proteins of about 7 to 9 Kd and evolutionary related [2]. Both seem to contain two transmembrane regions. Cyanobacteria seem to encode only for PSI-K.

As a signature pattern, the best-conserved region was selected which seems to correspond to the second transmembrane region.

-Consensus pattern: [GT]-F-x-[LIVM SEQ ID NO:4)]-x-[DEA]-x(2)-[GA]-x-[GTA]-[SA]-x-G-H-x-[LIVM SEQ ID NO:4)]-[GA]

[1] Golbeck J.H. Biochim. Biophys. Acta 895:167-204(1987).

[2] Kjaerulff S., Andersen B., Nielsen V.S., Moller B.L., Okkels J.S. J. Biol. Chem. 268:18912-18916(1993).

426. PTR2 family proton/oligopeptide symporters signatures

A family of eukaryotic and prokaryotic proteins that seem to be mainly involved in the intake of small peptides with the concomitant uptake of a proton has been recently characterized [1,2]. Proteins that belong to this family are: - Fungal peptide transporter PTR2.

- Mammalian intestine proton-dependent oligopeptide transporter PeptT1.

- Mammalian kidney proton-dependent oligopeptide transporter PeptT2.

- Drosophila opt1.

- Arabidopsis thaliana peptide transporters PTR2-A and PTR2-B (also known as the histidine transporting protein NTR1).

- Arabidopsis thaliana proton-dependent nitrate/chlorate transporter CHL1.

- Lactococcus proton-dependent di- and tri-peptide transporter dtpT.

- Caenorhabditis elegans hypothetical protein C06G8.2.

- Caenorhabditis elegans hypothetical protein F56F4.5.

- Caenorhabditis elegans hypothetical protein K04E7.2.

- Escherichia coli hypothetical protein ybgH.

- Escherichia coli hypothetical protein ydgR.

- Escherichia coli hypothetical protein yhiP.

- Escherichia coli hypothetical protein yjdL.

- Bacillus subtilis hypothetical protein yclF.

These integral membrane proteins are predicted to comprise twelve transmembrane regions. As signature patterns, two of the best